http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (19): 5226-5230, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5226.5230

A Novel Approach of Feature Selection Based on Decision-theoretic Rough Set Model

Qingfeng Duan
School of Economic and Management,
North University of China, Taiyuan, Shanxi, 030051, China

Abstract: Decision-theoretic rough set model is applied to the feature selection with capability of error tolerance which could deal with the decision problem with missing value or noise. Thus, a novel cost sensitive approach of feature selection is proposed. It tends to find appropriate reduct by the criterion of minimum cost computed by the theory of three-way Bayesian decisions and takes a reasonable selection with a heuristic strategy based on mutual information theory. To validate the proposed approach, classification performance comparison is employed on eight UCI datasets empirically. Consequently, it is demonstrated that the proposed approach work well and outperforms the others chosen in the experiments at most cases.

Key words: Feature selection, decision-theoretic rough set, cost sensitive

INTRODUCTION

To diminish the complexity of problem, feature selection has been a key procedure in resolving the problem of "curse of dimensionality" in the field of machine learning, pattern recognition. Many algorithms within artificial intelligence literature deal with feature selection. These algorithms can be classified in two categories including wrapper and filter. Wrapper methods treat performance algorithm itself as an evaluation function to estimate the accuracy of attribute subset. Thus, wrapper tends to be computationally expensive for the learning algorithm would be called repeatedly. Alternatively, filter methods discard undesirable attributes without regard to learning algorithm. Therefore, filter methods have been proven to be much faster than wrapper methods and they can be applied efficiently to large dataset with many attributes (Chen *et al*., 2011). The simple filter scheme is to evaluate each attribute individually, measuring its correlation to the target function (e.g., using a mutual information measurement) and then choose k attributes with the highest value.

Rough set model proposed by Pawlak (1982) has recently been applied to feature selection in the fields of machine learning, knowledge discovery, pattern recognition and so on (Liang *et al*., 2012). It provides a mathematical tool to discover data dependence and reduce the number of features contained in a dataset by purely structural methods. The complete solution is to generate all possible reduct and choose one satisfying a kind of criterion such as retaining the discriminative capability. Since it had been proved a NP-hard problem, incomplete solutions s earching for appropriate reduct are mainly focused on. For an instance, the heuristic algorithms employ an incremental strategy of hill-climbing algorithms to select appropriate features. The hill-climbing algorithms usually treat feature significance as heuristics to determine the search direction to reduct. Miao and Hou (2004) developed a mutual-information based rough set reduct approach. Inuiguchi *et al*. (2009) proposed a variable precision dominance-based rough set reduct approach. Qian *et al*. (2011) Introduced a theoretic framework based on rough set theory which is called positive approximation and can be used to accelerate a heuristic process for feature selection from incomplete data.

In addition, related techniques originated from artificial intelligence can also been employed to extend the discrimination capability of feature selection based on rough set model in some more sophisticated approach. Salamo and Lopez-Sanchez (2011) investigate feature selection based on rough sets for dimensionality reduction in Case-Based Reasoning classifiers. Derrac *et al*. (2012) proposed a global process of instance selection, carried out by a steady-state genetic algorithm, is combined with a fuzzy rough set based feature selection process which searches for the most interesting features to enhance both the evolutionary search process and the final preprocessed data set.

DECISION-THEORETIC ROUGH SET

A decision table is the following tuple:

$$S = (U, At = C D, V_a, f)$$

where, U is a finite nonempty set of objects, At is a finite nonempty set of attributes, C is a set of condition attributes describing the objects and D is a set of decision attributes that indicates the classes of objects. V_a is a nonempty set of values of a At and f: $U \rightarrow V_a$ is an information function that maps an object in U to exactly one value in V_a. An equivalence relation can be defined as $E \subseteq U \times U$, representing relationships between objects in U. Furthermore, an equivalence relation could be practically induced by a set of attributes such that the two objects with the equivalence relation would have the same values on each attribute. The equivalence class containing an object x is given by $[x]_E = \{y|y \in U, y\ E\ x\}$, or simply [x] if E is understood. Suppose $T \subseteq U$ is a set of objects representing an extension of concept. An important aim of rough set model is to describe T in term of equivalence class induced by conditional attributes.

With Pawlak rough set model, the decisions of acceptance or rejection are made without any error. In the practical classification, the classic model is too rigid to prefer making decisions of acceptance or rejection with the tolerance of error. For this purpose, some treatments with probability ideas are used to extend the Pawlak rough set model. Suppose P (T|[x]) is the conditional probability of an object belonging to T given that the object belongs to [x]. This probability can be estimated as P (T|[x]) = |T∩[x]|/|[x]|, where |•| denote the cardinality of a set. Whenever one choose a pair of thresholds (α, β) with α>β, three probabilistic regions are introduced as:

$$\begin{aligned} &POS(T) = \{x \mid x \in U, P(T \mid [x]) \geq \alpha\}, \\ &BND(T) = \{x \mid x \in U, \alpha < P(T \mid [x]) < \beta\}, \\ &NEG(T) = \{x \mid x \in U, P(T \mid [x]) \leq \beta\} \end{aligned} \quad (1)$$

According to making decision of three probabilistic regions, the cost or risk of misjudgment can be measured in term of Bayesian techniques. Let $\Omega = \{\Omega_1, \ldots, \Omega_s\}$ be a finite set of s states and $A = \{a_1, \ldots a_m\}$ be a finite set of m possible actions. Let $\lambda(a_i|\Omega_j)$ denote the loss or cost for taking action a_i when the state is Ω_j. Suppose action a_i is taken now, the expected cost associated with taking action a_i is introduced by:

$$R(a_i \mid x) = \sum_{j=1}^{s} \lambda(a_i \mid \omega_j) \cdot P(\omega_j \mid x) \quad (2)$$

In decision-theoretic rough set models, the set of states is $\Omega = \{T, T^C\}$, indicating an object is in a decision class T or not. The set of actions is $A = \{a_P, a_B, a_N\}$, where, a_P, a_B and a_N represent the three actions in classifying x, namely, deciding x∈POS (T), deciding x∈BND (T) and deciding x∈NEG (T) respectively. Let λ_{PP}, λ_{BP}, λ_{NP} denote the cost caused by taking actions a_P, a_B a_N respectively when an object belongs to T. and λ_{PN}, λ_{BN}, λ_{NN} denote the cost caused by taking actions a_P, a_B a_N respectively when an object does not belongs to T. Given the parameters of Bayesian decision cost, the expected cost according to make different decisions for object x in [x] is presented as:

$$\begin{aligned} R_P &= R(a_P|[x] = \lambda_{PP} \cdot P(T|[x]) + \lambda_{PN} \cdot (PT^C|[x]) \\ R_B &= R(a_B|[x]) = \lambda_{BP} \cdot P(T|[x]) + \lambda_{BN} \cdot P(T^C|[x] \\ RN &= R(a_N|[x]\_ = \lambda_{NP} \cdot) P(T|[x]) + \lambda_{NN} \cdot P(T^C|[x]) \end{aligned} \quad (3)$$

These parameters in cost function are obviously critical to the Bayesian decision and should be considerably given appropriate values. In common sense, there exists relationships with $\lambda_{PP} < \lambda_{BP} < \lambda_{NP}$ and $\lambda_{PN} < \lambda_{BN} < \lambda_{NN}$ according the definition of decisions cost. In the three-way decisions proposed by Yao (2010), the parameters α and β in the Eq. 1 could be determined by the cost function λ respectively as following:

$$\begin{aligned} \alpha &= \frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{PN} - \lambda_{BN} + \lambda_{BP} - \lambda_{PP}} \\ \beta &= \frac{\lambda_{BN} - \lambda_{NN}}{\lambda_{BN} - \lambda_{NN} + \lambda_{NP} - \lambda_{BP}} \end{aligned} \quad (4)$$

In the decision table S, one can divide the university U into three regions of partition π_D by considering the relationship with partition π_A, based on threshold α and β.

$$\begin{aligned} POS_{(\alpha,\beta)}(\pi_D|\pi_A) &= \{x \in U | P(D_{max}([x]_A)|[x]_A) \geq \alpha\} \\ BND_{(\alpha,\beta)}(\pi_D|\pi_A) &= \{x \in U | \alpha < P(D_{max}([x]_A)|[x]_A < \beta \\ NEG_{(\alpha,\beta)}(\pi_D|\pi_A) &= \{x \in U | P(D_{max}([x])_A)|[x]_A) \leq \beta\} \end{aligned} \quad (5)$$

Where:

$$D_{max}([x]_A) = \arg\max D_i \in \pi_D([x]_A \cap D_i)$$

FEATURE SELECTION ALGORITHM

Reduction by removing the irrelevant features can effectively reduce the dimensions of problem. For most feature selection algorithms based on rough set models, the discernible capability of selected features is regarded as the critical criterion in judging the reduction. According to Pawlak rough set model, it is reasonable that the selected features, a subset of conditional attributes in decision table, hold the same discernible capability with all of features, by making the positive region of objects unchanged after removing some redundant features.

Decision-theoretic rough set model mentioned before extend the framework of Pawlak's model to deal with the probability decision problems by allowing certain acceptable level of errors. Yao (2010) proposed the three-way decision by using the techniques of Bayesian theory. More specifically, rules generated by the three regions, illustrated in the Eq. 5, form three-way decision rules: The positive rules generated by positive region make decisions of acceptance; the negative rules generated by negative region make decisions of rejection; the boundary rules generated by boundary region make decisions of deterred or no-committed decisions. Cost or risk of making decisions based on the three-way rules would inevitably exist in the procedure of Bayesian decisions.

By reviewing the definition of decision-theoretic rough set, it is reasonable to regarding the minimum cost of making decisions as a critical criterion when judging the performance of features selections. The Bayesian decision procedure deals with the making decisions with the minimum cost based on the observed evidence. A change of three regions after taking away some attributes may result in an increment of decision cost. Therefore, the solution of feature selection in the decision-theoretic rough set is to find a subset of attributes while still holding the same minimum decision cost as all of attributes. In other words, the minimum cost should be achieved under the condition of selected features. Given attribute set $A \subseteq C$, decisions cost with A is defined as the sum of all cost associated with decisions allocating each object into one of three regions. Then the total cost equation can be presented as:

$$R_A = \sum_{x_i \in POS_{(\alpha,\beta)}(\pi_D|\pi_A)} R_P + \sum_{x_i \in BND_{(\alpha,\beta)}(\pi_D|\pi_A)} R_B \sum_{x_i \in NEG_{(\alpha,\beta)}(\pi_D|\pi_A)} R_N \quad (6)$$

The objective of feature selection procedure is the optimization solution finding the subset of C with the minimum cost. Then reduction here can be defined as: A is minimum cost attribute reduct if and only if (1) $A = \arg\min_{A \subseteq C} (R_A)$, (2) $\forall A' \subset A$, $R_A < R_{A'}$. Instead of obeying the rule keeping the positive region unchangeable, the minimum cost of Bayesian decisions by the way of three-way is alternatively regarded as the key criterion to select features.

According to Shannon's information theory, entropy is a key measurement of information. Since it is capable of quantity the uncertainty of random variables and scaling the amount of information shared by them effectively, it has been widely used in many fields. Uncertainty of a random variable X with discrete value can be measured by entropy H (X) as:

$$H(X) = -\sum_{x \in X} p(x)\log_2 p(x) \quad (7)$$

where, p (•) is the probability mass function of variable X. Given a random variable Y also with discrete value, the conditional entropy H (X|Y) is expressed as:

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y)\log_2 p(x|y) \quad (8)$$

where, p (x|y) is the conditional probability mass of X and Y. Conditional entropy refers to the uncertainty reduction of variables when other is known. To quantity how much information is shared by two variables, a concept termed mutual information I (X; Y) is defined as:

$$I(X; Y) = H(X)-H(X|Y) \quad (9)$$

Computation of finding optimal solution of features with minimum cost in making decision is NP hard. Instead of exhaustion algorithm, many heuristic approaches have been investigated to solve the optimal problem of attribute reduct using rough set theory in linear time. Considering the differences among features in capability of discerning the equivalent class associated with decision attribute, features with the higher or the highest discernible power should be possessed the priority in the procedure of searching potential features. Hence, mutual information is used to give the heuristic idea for finding the optimal solution in the procedure of feature selection.

Features can be typically selected in three ways, i.e., forward, backward and random. In this paper, feature selection procedure is considered in a straight forward. At each iteration, someone attribute from candidates is selected according to information theoretic based criterion computed by mutual information theory which indicates its contribution to the discernible capability. Namely, candidate attributes are ranked in descending order by their mutual information with respect to decision attribute and the highest ranked candidate is given priority to be selected. The selection iterations can conduct until the target of minimum cost of decision is achieved. Thus, the heuristic algorithm based on minimum cost is illustrated as following:

```
Input: The decision table.
Output: The reduct A
Begin
    A = ø, G = C;
    Compute I (a;D) for all a?C;
    While RA>RC and G?ø;
        Select an attribute in term of I (a;D) ;
        A = A?{a};
        G = G-{a};
    End While
    output A;
End Begin
```

EXPERIMENTS AND RESULTS

To test the proposed method empirically, eight real world datasets from UCI's machine learning data repository are adopted in the simulation experiments. These datasets contain various numbers of features and come from different domains. Considering some missing values arisen from various aspects as stated in the online documentations for these datasets, each missing values are replaced by the mean for numeric features and the mode for nominal ones.

In addition to feature selection algorithms, four representative classifiers are employed, i.e., C4.5, 1-Nearest neighbor, SVM and Naïve Bayes which are the most influential algorithms that have been widely used in data mining community. The experimental workbench is Weka (Waikato environment for knowledge analysis) which is a collection of machine learning algorithms for data mining tasks. The parameters of classifiers for each experiment are set to default value of Weka. For estimating the performance of classification algorithms, 10-fold cross validation is used.

We empirically evaluate the performance of the proposed method (Alg. 1) by comparing with two typical feature selectors: Alg. 2 and Alg. 3. Alg. 2 (Jia *et al*., 2013) is also a cost sensitive approach considering the cost three-way Bayesian decision as a criterion for feature selection and adopts a heuristic approach with a cost based fitness function to find the solution. To test the performance of cost sensitive feature selection compared with the positive region based reduct approach, Alg. 3 which is a classic algorithm proposed by Pawlak in rough set model, is also carried out in the experiments. Alg. 3 tends to find a subset of attributes with the same positive regions as the whole attributes such that the original discriminative capability could be hold by less features. Parameters of decisions cost in the proposed algorithm are cautiously set for getting a satisfied classification performance. Practically, λ_{PP} and λ_{NN} are zero because of no cost arisen by the right judgments and λ_{BP}, λ_{BN}, λ_{NP}, λ_{PN} are 0.2, 0.2, 0.8, 0.8 respectively.

The experimental results about classification accuracies on eight datasets for four chosen classifiers using three feature selection algorithms are present in Table 1. The bold value in entries means that it is the largest one among these three feature selection algorithms in the same classifier corresponding to each dataset. The average accuracies with the same selector are given in the "Ave." row. In order to investigate the efficiency of selectors with the same classifier, an indicator named average efficiency value is presented in the AVE row corresponding to the each experimental result table.

From the perspective of classification accuracies, Alg. 1 shows the better performance than others by employing the classifiers except the classifier on C4.5. Specially, when 1-Nearest Neighbor employed as classifier, Alg. 1 shows the better performance of accuracies than other reduct algorithms nearly in the all cases except the Glass dataset. For example, the numbers of cases for which Alg. 1 achieves significantly higher classification accuracies over Alg. 2 and Alg. 3 are eleven, one and three. For the classifiers of SVM and Naïve Bayes, one may also observe the similar fact that Alg. 1 clearly surpasses others in most cases. In the experiments of C4.5 as classifier, it is demonstrated that Alg. 3 gain the dominance in the classification accuracies and Alg. 2 exhibit the worst performance that only winning twice in the comparison of classification accuracies. And the average value of classification accuracies for the eight real datasets also denotes that proposed method perform better than others for these classifiers with the exception of C4.5. For different categories of classifiers, there exist varieties when ranking classification performance using three compared feature selection algorithms. Namely, feature selection based on three-way decision theoretic model exhibit better classification accuracies than that based on positive region invariance with SVM or Naïve Bayes employed.

Table 1: Comparison of classification accuracies of four classifiers

	C4.5			1-Nearest-Neighbor			SVM			Naïve Bayes		
	Alg. 1	Alg. 2	Alg. 3	Alg. 1	Alg. 2	Alg. 3	Alg. 1	Alg. 2	Alg. 3	Alg. 1	Alg. 2	Alg. 3
Car	0.927	0.786	0.927	0.938	0.786	0.938	0.930	0.811	0.930	0.847	0.713	0.847
Breast	0.747	0.715	0.715	0.695	0.687	0.600	0.667	0.687	0.687	0.715	0.716	0.716
Glass	0.562	0.562	0.564	0.597	0.597	0.599	0.679	0.679	0.657	0.598	0.598	0.564
Bridges	0.552	0.552	0.574	0.638	0.638	0.526	0.733	0.733	0.471	0.816	0.816	0.481
Credit	0.854	0.854	0.854	0.824	0.818	0.824	0.865	0.862	0.865	0.849	0.850	0.849
Syobean	0.923	0.964	0.800	0.816	0.78	0.752	0.934	0.941	0.928	0.871	0.799	0.852
Bands	0.601	0.689	0.752	0.830	0.662	0.705	0.803	0.660	0.681	0.768	0.660	0.704
Synthetic	0.760	0.632	0.751	0.764	0.660	0.741	0.803	0.775	0.79	0.803	0.779	0.785
Ave.	0.741	0.719	0.742	0.763	0.704	0.711	0.802	0.769	0.751	0.783	0.741	0.725
AEV	0.114	0.114	0.109	0.114	0.112	0.104	0.122	0.121	0.111	0.119	0.112	0.106

CONCLUSION

In this study, a novel cost sensitive approach is proposed. With difference to classic positive region based feature selection approach, the cost sensitive approach considers the three-way decision cost as the critical criterion for attribute reduct based on the decision-theoretic rough set model. For the error tolerance characteristics of decision-theoretic model, the proposed feature selection approach can deal with the inconsistence dataset in the real world better than the classic reduct approach based on rough set model. To verify the effectiveness of the proposed algorithm, empirical experiments are employed on the eight datasets from UCI. Consequently, it is demonstrated that the proposed approach with feature selection strategy based on mutual information theory work well and outperform the others chosen in the experiments at most cases.

ACKNOWLEDGMENT

The research has been supported by the project of Science Foundation of NUC in 2013.

REFERENCES

Chen, Y., D. Miao, R. Wang and K. Wu, 2011. A rough set approach to feature selection based on power set tree. Knowledge-Based Syst., 24: 275-281.

Derrac, J., C. Cornelis, S. Garcia and F. Herrera, 2012. Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection. Inform. Sci., 186: 73-92.

Inuiguchi, M., Y. Yoshioka and Y. Kusunoki, 2009. Variable-precision Dominance-based rough set approach and attribute reduction. Int. J. Approx. Reason., 50: 1199-1214.

Jia, X., W. Liao, Z. Tang and L. Shang, 2013. Minimum cost attribute reduction in decision-theoretic rough set models. Inform. Sci., 219: 151-167.

Liang, J., F. Wang, C. Dang and Y. Qian, 2012. An efficient rough feature selection algorithm with a multi-granulation view. Int. J. Approximate Reasoning, 53: 912-926.

Miao, D. and L. Hou, 2004. A comparison of rough set methods and representative inductive learning algorithms. Fundamenta Inform., 59: 203-219.

Pawlak, Z., 1982. Rough sets. Int. J. Comput. Inform. Sci., 11: 341-356.

Qian, Y., J. Liang, W. Pedrycz and C. Dang, 2011. An efficient accelerator for attribute reduction from incomplete data in rough set framework. Pattern Recognit., 44: 1658-1670.

Salamo, M. and M. Lopez-Sanchez, 2011. Rough set based approaches to feature selection for Case-based reasoning classifiers. Pattern Recognit. Lett., 32: 280-292.

Yao, Y., 2010. Three-way decisions with probabilistic rough sets. Inform. Sci., 180: 341-353.