Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 38 (2011) 4443-4446

Awakening to reality

Elixir ISSN: 2229-712X

Comparative analysis of various feature selection algorithms based on fuzzy-rough set approach

V. Abirami Nachammai[1], A.Pethalakshmi[2]

[1]Thiagarajar College of Engineering, Madurai, TamilNadu, India.

[2]Department of Computer Science, M.V.M. Govt. Arts College for Women, Dindigul, TamilNadu, India.

ARTICLE INFO

Article history:

Received: 22 August 2011;
Received in revised form:
26 August 2011;
Accepted: 31 August 2011;

Keywords

Data mining,
Improved Quickreduct,
Quickreduct,
Rough Set.

ABSTRACT

Rough Set Theory provides a formal framework for data mining. Feature Selection or Attribute Reduction is a preprocessing step in data mining, and it is very effective in reducing dimensionality, reducing irrelevant data, increasing learning accuracy and improving comprehensibility. The fuzzy-rough feature selection algorithm was used to handle the continuous real-valued data as well as to handle noisy data. It was implemented by standard fuzzification techniques enabling linguistic labels to be associated with attribute values. It also provides uncertainty modeling by allowing the possibility of the membership value to more than one fuzzy label. In this paper, we use an Improved Quickreduct algorithm by redefining the lower and upper approximations based on fuzzy set theory. The membership degrees of feature values to fuzzy sets are exploited in the process of dimensionality reduction. The experiments are carried out on public domain datasets available in UCI machine learning repository and real Tuberculosis data set.

Introduction

Data Mining

Data mining refers to extracting or "mining" knowledge (hidden information) from large amounts of data. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data pattern analysis, data archaeology and data dredging. Data mining treats as synonym for another popularly used term, Knowledge Discovery in Databases (KDD) [3]. KDD consists of the following steps to process it such as Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation and Knowledge presentation.

KDD is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining is not a single technique, some commonly used techniques are: Statistical Methods, Case-Based Reasoning (CBR), Neural Networks, Decision Trees, Rule Induction, Bayesian Belief Networks (BBN), Genetic Algorithms, Fuzzy Sets and Rough Sets.

Rough Set Theory

Rough Set Theory was initially developed [9, 10] for a finite universe of discourse in which the knowledge base is a partition, which is obtained by any equivalence relation defined on the universe of discourse. In rough sets theory, the data is organized in a table called decision table. Rows of the decision table correspond to objects, and columns correspond to attributes. In the data set, a class label to indicate the class to which each row belongs. The class label is called as decision attribute, the rest of the attributes are the condition attributes. Here, C is used to denote the condition attributes, D for decision attributes, where C □ D = Φ, and tj denotes the jth tuple of the data table. Rough sets theory defines three regions based on the equivalent classes induced by the attribute values: lower approximation, upper approximation, and boundary. Lower approximation contains all the objects, which are classified surely based on the data collected, and Upper approximation contains all the objects, which can be classified probably, while the boundary is the difference between the upper approximation and the lower approximation. Hu et al., [4] presented the formal definitions of rough set theory. Kusiak [8] described the basic concepts of rough set theory, and other aspects of data mining.

Let U be any finite universe of discourse. Let R be any equivalence relation defined on U. Clearly, the equivalence relation partitions U. Here, (U, R) which is the collection of all equivalence classes, is called the approximation space. Let W_1, W_2, W_3, …, W_n be the elements of the approximation space (U, R). This collection is known as knowledge base. Then for any subset A of U, the lower and upper approximations are defined as follows:

$$\underline{R}A = \cup\{W_i / W_i \subseteq A\}$$

$$\overline{R}A = \cup\{W_i / W_i \cap A \neq \varnothing\}$$

The ordered pair ($\underline{R}A$, RA) is called a rough set. Once defined these approximations of A, the reference universe U is divided into three different regions: the positive region $POS_R(A)$, the negative region $NEG_R(A)$ and the boundary region $BND_R(A)$, defined as follows:

$$POS_R(A) = \underline{R}A$$

$$NEG_R(A) = U - \overline{R}A$$

$$BND_R(A) = \overline{R}A - \underline{R}A$$

Hence, it is trivial that if BND(A) = Φ, then A is exact. This approach provides a mathematical tool that can be used to find out all possible reduces.

Feature Selection

Feature selection process refers to choose a subset of attributes from the set of original attributes. The purpose of the feature selection is to identify the significant features, eliminate the irrelevant of dispensable features to the learning task, and

Tele:
E-mail addresses: abiramicse@tce.edu

build a good learning model. The benefits of feature selection are twofold: it considerably decreased the computation time of the induction algorithm and increased the accuracy of the resulting mode.

The paper is organized as follows: Section 2 briefs about the data sets used for this study. Quickreduct algorithm is described in Section 3. Section 4 deals about Fuzzy-Rough sets. Comparative analysis of Quickreduct and Improved Quickreuct is described in Section 5. Section 6 states the conclusion of this paper.

Data preparation

The medical data sets viz., Postoperative, Pima, New-Thyroid obtained from UCI machine learning repository [1] and the real Tuberculosis data set are considered for this study. The Tuberculosis data set is collected from Government Hospital, Kodaikanal, Dindigul District, Tamilnadu. The advantage of this dataset is that it includes a sufficient number of records of different types. The set of descriptors represents all the potentially interesting typically available information about the Tuberculosis. The record of every patient contains 17 condition attributes and one decision attribute. The details of attributes are given as follows: The condition attributes are Gender, Age, Cough, Coughing up Blood, Chest Pain, Fever, Swollen Glands, Stomach Pain, Loss of Appetite, Weight Loss, Fatigue, Head Ache, Vomiting, Chills, Night Sweat, Difficulty in Breathing, Low Blood Pressure and the Decision attribute Result.

Quickreduct Algorithm(QR)

The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature as the original. A reduct is defined as a subset of minimal cardinality R_{min} of the conditional attribute set C such that $\gamma_R(D) = \gamma_C(D)$.

$$R = \{X : X \subseteq C;\ \gamma_x(D) = \gamma_C(D)\}$$

$$R_{min} = \{X : X \in R;\ \forall Y \in R;\ |X| \leq |Y|\ \}$$

The intersection of all the sets in R_{min} is called the core, the elements of which are those attributes that cannot be eliminated without introducing more contradictions to the dataset. In this method a subset with minimum cardinality is searched for. The pseudocode of the Quickreduct algorithm [5, 6, 7] is given below:

Quickreduct(C,D)

```
C, the set of all conditional features;
D, the set of decision features.
(a) R ← {}
(b) Do
(c) T ← R
(d) ∀ x ∈ (C-R)
(e) if $\gamma_{R\cup\{x\}}(D) > \gamma_T(D)$
        where $\gamma_R(D)$=card($POS_R(D)$) / card(U)
(f) T ← R∪{x}
(g) R ← T
(h) until $\gamma_R(D) == \gamma_C(D)$
(i) return R
```

Fuzzy-Rough Sets

In [2], the fuzzy p-lower and p-upper approximations are defined as:

$$\underline{\mu}_{pX}(F_i) = \inf_x \max\{1-\mu_{Fi}(x),\ \mu_X(x)\}\ \forall i \qquad (4.1)$$

$$\overline{\mu}_{pX}(F_i) = \sup_x \min\{\mu_{Fi}(x),\ \mu_X(x)\}\ \forall i \qquad (4.2)$$

where F_i denotes a fuzzy equivalence class belonging to U / P. As the universe of discourse in feature selection is finite, the use of sup and inf are to be altered.

As a result of this, the fuzzy lower and upper approximations are herein redefined as:

$$\underline{p}X(x) = \max(F(x),\ (\inf_{y\in U} \max\{1 - F(y),\ X(y)\}) \qquad (4.3)$$

$$\overline{p}X(x) = \max(F(x),(\sup_{y\in U} \min\{F(y),\ X(y)\}) \qquad (4.4)$$

Here F(i) denotes the normalized object values. Using the lower approximation in (4.3), the positive region can be defined as

$$POS_{p(Q)}(x) = \sup_{X\in U/Q} \underline{p}X(x) \qquad (4.5)$$

Using the definition of the positive region, the new dependency function can be defined as follows:

$$\gamma_p(Q) = |POS_{p(Q)}(x)| / |U|$$
$$= \sum_{x\in U} POS_{p(Q)}(x) / |U| \qquad (4.6)$$

Improved Quickreduct Algorithm (IQR)

The Quickreduct algorithm is improved herein under using the lower and upper approximations in (4.3) and (4.4) in order to obtain the best degree of dependency value by normalizing the information system. That is, the maximum value is selected and all other values are divided by the maximum value [11]. Then the Improved Quickreduct Algorithm is applied to obtain the minimal reduct. The Improved Quickreduct Algorithm is as follows :

Improved Quickreduct(C, D)

```
C, the set of all conditional features
Q, the set of all decision features
(a) C← the set of all normalized values
(b) RED ← {}
(c) Do
(d) TEMP ← RED
(e) $\gamma_{best} = 0$
(f) For x ∈ C
(g) if $\gamma_{RED\cup(x)}(Q) > \gamma_{best}(Q)$ (using 4.3, 4.5 and 4.6)
(h) TEMP ← RED ∪ (x)
(i) $\gamma_{best}$ ← TEMP
(j) RED ← TEMP
(k) until $\gamma_{best} == \gamma_C(Q)$
(l) return RED
```

Worked Example

A system of 8 data points consisting four condition attributes and a decision attribute, adopted from [3] is taken into consideration and it is presented in Table 4.1.

The following substitutions Low = 1, Medium = 2, High = 3, Com = 1 and Sub = 2 can be used. The normalized values are tabulated in Table 4.2

In Table 4.2 {Weight} is assigned to A, {Door} is assigned to B, {Size} is assigned to C, {Cylinder} is assigned to D and {Mileage} is assigned to Q. The lower approximations of A, B, C and D are calculated using (5.4). The method of computing the lower approximation of the attribute A is elaborated here. For a class X = {1, 3, 6} in decision attribute, A{1, 3, 6}(x) needs to be calculated in order to compute the decision equivalence class.

$$\underline{A}_{\{1,3,6\}}(x) = \max(F(x),\ (\inf_{y\in U} \max\{1 - F(y),\ X(y)\})$$

For object 1, this can be calculated as follows:

max (1–a(1), X(1)) = max(0.8333, 1.0)
= 1.0

max (1– a(2), X(2)) = max(0.8333, 0.0)
= 0.8333

max (1 – a(3), X(3)) = max(0.6667, 1.0)

$= 1.0$

$\max(1 - a(4),\ X(4)) = \max(0.6667,\ 0.0)$

$= 0.6667$

$\max(1 - a(5),\ X(5)) = \max(0.5000,\ 0.0)$

$= 0.5000$

$\max(1 - a(6),\ X(6)) = \max(0.8333,\ 1.0)$

$= 1.0$

$\max(1 - a(7),\ X(7)) = \max(0.5000,\ 0.0)$

$= 0.5000$

$\max(1 - a(8),\ X(8)) = \max(0.8333,\ 0.0)$

$= 0.8333$

Therefore, max (A(x))

$= \max(0.1667,\ \inf\{1.0,\ 0.8333,\ 1.0,\ 0.6667,\ 0.5000,\ 1.0,\ 0.5000,\ 0.8333\})$

$= \max(0.1667,\ 0.5000)$

$= 0.5000$

Thus, $\underline{A}_{\{1,\ 3,\ 6\}}(1) = 0.5000$. Similarly, $\underline{A}_{\{1,\ 3,\ 6\}}$ can be computed for the other objects. Then the corresponding class values for X = {2, 4, 5, 7, 8} can also be determined. For object1, $\underline{A}_{\{2,\ 4,\ 5,\ 7,8\}}(1) = 0.6667$. Similarly, $\underline{A}_{\{2,\ 4,\ 5,\ 7,\ 8\}}$ can be computed for the other objects. Using these values, the positive region for each object can be calculated using

$$POS_{p(Q)}(x) = \sup_{X\in U/Q} \underline{A}X(x)$$

For object1,

$POS_{p(Q)}(1) = \sup(0.5000,\ 0.6667) = 0.6667.$

Similarly, $POS_p(Q)$ can be computed for the other objects. The next step is to determine the degree of dependency of Q on A:

$\gamma_A(Q) = 5.3336/\ 8 = 0.6667$

Similarly B, C and D are calculated to get the following degree of dependency.

$\gamma_B(Q) = 0.5417$

$\gamma_C(Q) = 0.8333$

$\gamma_D(Q) = 0.8333$

The attribute 'A' is chosen and added to the potential reduct. The max (A, B, C, D) should be calculated from the Table 4.2 to compute $\gamma_{\{A,\ B,\ C,\ D\}}(Q)$ which yields

$\gamma_{\{A,\ B,\ C,\ D\}}(Q) = 0.8333$

and then it has to be compared with $\gamma_A(Q)$. The reduct attributes can be obtained only if the values are equal.

Otherwise this attribute value should be combined with the next attribute value. That is, take the max(A, B) from the Table 4.2. This process iterates and the two dependency degrees are calculated as,

$\gamma_{\{A,\ B\}}(Q) = 0.6667$

This value is equal to $\gamma_A(Q)$, and so "if" condition is not satisfied, hence the next combination of max(A, C) should be taken and the dependency degrees are calculated as

$\gamma_{\ (A,\ C)}(Q) = 0.8333$

This value is greater than $\gamma_A(Q)$, and it is added to the potential reduct and the "until" condition is now satisfied, and the final output of reduct attributes {Weight, Size} for car data set is obtained.

Comparative Analysis

The Quickreduct and the Improved Quickreduct algorithm have been implemented for medical databases available in the UCI data repository and the real Tuberculosis data. However, the Quickreduct algorithm is not guaranteed to find a minimal subset. The fuzzy lower and upper approximations are used to improve the Quickreduct algorithm. The comparative analysis of Quickreduct algorithm and the Improved Quickreduct Algorithm is tabulated in Table 5.1

Reduced attributes obtained for Tuberculosis data set after applying Improved Quickreduct algorithm are: Cough, Coughing up Blood, Chest Pain, Fever, Swollen Glands, Loss of Appetite, Weight Loss, Fatigue, Vomiting, Difficulty in Breathing.

Conclusion

This paper was focused on a sub-problem encountered in a data mining namely, feature selection or attribute reduction via Quickreduct and Improved Quickreduct algorithm. It was evident that the Improved Quickreduct algorithm produced minimal reduct for all the data sets. In our Tuberculosis data set, among 17 attributes only 10 attributes had been chosen for decision making. It was proved that the 10 attributes alone are sufficient enough for diagnosing Tuberculosis patient or not. This investigation not only helps in saving the computing time and memory space, but enables future research oriented data selection also.

References

[1] C. L. Blake and C. J. Merz, UCI Repository of machine learning databases, Irvine, University of California, http://www.ics.uci.edu/~mlearn/. 1998.

[2] D. Dubois and H. Prade, "Putting Rough Sets and Fuzzy Sets together". In: R. Slowinski (Ed.) Inelligent Decision Support, Kluwer Academic Publishers, pp: 203–232, 1992.

[3] J. Han and M. Kamber, Data mining: Concepts and Techniques, Morgan Kaufmann Publishers, 1995.

[4] X. Hu, T.Y. Lin and J. Jianchao, A New Rough Sets Model Based on Database Systems, Fundamenta Informaticae, 1-18, 2004.

[5] R. Jensen and Q. Shen, Fuzzy-rough attribute reduction with application to web categorization, Fuzzy Sets and Systems, Vol.141, No.3, 469-485, 2004.

[6] R. Jensen and Q. Shen, Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No.12, 2004.

[7] R. Jensen, Combining Rough and Fuzzy Sets for Feature Selection, Ph.D Thesis, School of Informatics, University of Edinburgh, 2005.

[8] A. Kusiak, Rough Set Theory: A Data mining Tool for Semiconductor Manufacturing, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24, No.1, 2001.

[9] Z. Pawlak, Rough Sets, International Journal of Computer and Information Sciences, Vol.11, No.5, 341-356, 1982.

[10] Z. Pawlak, Rough Sets:Theoritical Aspects and Reasoning about Data, Kluwer Academic Publishers, 1991.

[11] T. J. Ross, "Fuzzy Logic with Engineering Applications",McGraw-Hill Inc., 1995.

Table 4.1. Car Data Set

Object	Weight	Door	Size	Cylinder	Mileage
1	Low	2	Com	4	High
2	Low	4	Sub	6	Low
3	Medium	4	Com	4	High
4	High	2	Com	6	Low
5	High	4	Com	4	Low
6	Low	4	Com	4	High
7	High	4	Sub	6	Low
8	Low	2	Sub	6	Low

Table 4.2. Normalized Values

Object	Weight	Door	Size	Cylinder	Mileage	
					(1,3,6)	(2,4,5,7, 8)
1	0.1667	0.3333	0.1667	0.6667	1	0
2	0.1667	0.6667	0.3333	1.0000	0	1
3	0.3333	0.6667	0.1667	0.6667	1	0
4	0.5000	0.3333	0.1667	1.0000	0	1
5	0.5000	0.6667	0.1667	0.6667	0	1
6	0.1667	0.6667	0.1667	0.6667	1	0
7	0.5000	0.6667	0.3333	1.0000	0	1
8	0.1667	0.3333	0.3333	1.0000	0	1

Table 5.1. Comparative Analysis of Quickreduct and Improved Quickreduct

Data set	Instances	No. of Attributes	QR	IQR
Postoperative	90	8	4	3
Pima	768	8	5	3
New-Thyroid	215	5	4	4
Tuberculosis	500	17	12	10