Rough set based feature selection using modified rough membership function

Prof. Dr. Ahmad T. Sadiq
Department of computer science / University of Technology
Baghdad/Iraq
Sura N. Maryoush
Computer science Dept./collage of education / Al-Mustansiriya university
Baghdad/Iraq

Abstract:

feature selection (FS) is one of the important steps in the knowledge discovery, which aims to reduce the dimensionality of data. In this paper, a feature selection algorithm is proposed. The proposed algorithm use the rough membership function, which is modified in order to be suitable for measuring the effectiveness of each attribute value, and then using it for measuring the effectiveness of each attribute through a new formula called a modified attribute membership (MAM). The experiments shows that the proposed algorithm provides an effective tool for selecting feature and reducing the dimensionality of data.

Keywords: feature selection, modified rough membership function, modified attribute membership, rough set.

1. Introduction

One of the important steps in the discovery of knowledge from data is a feature selection. The Feature selection process can be defined as an attempt to select a subset of features from the complete set of features that are more informative than unselected features, which have the ability to represent the dataset as using the complete set of features. Feature selection aims to reduce the dimensionality of the data without decreasing the accuracy of the prediction. Many important advantages are obtained from the feature selection algorithms for learning algorithms, such as reducing the computation time of the induction algorithms, improve the accuracy and decrease the complexity of the generated rules, and decreases the storage requirement and cost of computations. Also feature selection can enhance the performance of the pattern classifiers by improving the overall speed and quality of the pattern classification process [1,2,3].

Rough set theory was introduced by Zdzislaw Pawlak in 1982. It is one of the effective methods that concerned with analysis of data tables for the classification purposes, and dealing with uncertainty and vagueness that exist in the data tables [4]. Rough sets consider as an extension to the set theory, which uses approximations in

order to make decisions[5]. Rough set theory has been used in the development of many feature selection algorithms [6].

In this paper, a new feature selection algorithm is proposed. This feature selection algorithm is based on rough membership function. Rough membership function is modified to be suitable for measuring the information amount of each attribute value, then measuring the information amount of the whole attribute. The proposed algorithm is tested using five datasets, which contains different data from different sources.

The other sections in this paper are organized as follows: section 2 presents a literature related of rough set based feature selection algorithm. Section 3 explains the main concepts of rough set theory. Section 4 describes the proposed feature selection algorithm based on modified rough membership function. The experimental results and discussion are presented in section 5, followed by the conclusion in section 6.

2. Related works

Many literatures related of feature selection algorithms based on rough set are proposed. In 2011, L. Sun, et al., firstly discussed the limitations of the feature selection algorithms that exist recently. Secondly, a new rough entropy is proposed in order to measure the knowledge roughness, then a new significance measure of features depending on the rough entropy was presented. The new significance measure is more efficient when it compared with significance measures that are based on positive region and conditional information entropy [7] .

Also in 2011,C. Velayutham and K.thangavel proposed an unsupervised feature selection based on rough set, where the decision class labels are not required to be provided to select a subset of features since features are not required to be related to the decision classes. The unsupervised feature selection uses a relative dependency measure and employs a backward elimination search in order to remove features from the set of complete features. The proposed method proves its efficiency and its effectiveness in removing the redundant features [8] .

In 2013, K. Anitha and P. Venkatesan used quick reduct algorithm to reduce the data size and selecting a subset of features, and also discussed the basic applications and concepts of rough set in feature selection. Rough set in feature selection has the advantage of not requiring any additional parameters that differ from the original data. When the collected data are not precise, incomplete or redundant about objects of the domain, rough set can be a useful method for dealing with this kind of data [9] .

In 2014, T. Sridevi and A. Murugan proposed a modified correlation rough set as feature selection algorithm to predict both prognosis and diagnosis breast cancer, where the Wisconsin data sets of breast cancer are used. The proposed approach consists of two levels. In the first level of feature selection, feature selection process is based on rough set with using different starting reduct values. In the second level, the

reduced set which is resulted from the first level is used to select features from it, and the selection is based on the correlation feature selection. The proposed method is effective in performance of classification and in term of number of features that are selected [10] .

3. Rough sets concepts

As mentioned before, rough sets deals with analyzing the uncertainty and vagueness that may be found in data [11]. The vagueness concepts are approximated by rough set, where the rough set presents two precise concepts, named lower approximation and upper approximation, which represents a classification of the interest domain into disconnected categories. The elements of the domain that certainly belong to the interest subset are described by means of the lower approximation, while the elements that not certainly belong to the interest subset are described by means of the upper approximation [5].

The vagueness is expressed in rough sets by using boundary region of the set, in additional to the means of set elements membership. The boundary region consists of elements that exist in the upper approximation and not exist in the lower approximation (the difference between the upper and lower approximations). The set is counted as crisp set, if its boundary region is empty, otherwise it is considered as a rough set [12]. In real life, the data have different levels of complexity and sizes, which makes the data is difficult to be analyzed and also hard to be managed from computational view point. The main aims of Rough Set analysis are to handle inconsistency that exist in data and to reduce the size of data [11].

The basic concepts of rough set theory can be explained by the following definitions[4] :

Definition 3.1: let S be an information system, S=(U, A), where U represents a finite set of instances called a universe, and A represents a finite set of attributes. S is called a decision table if A is distinguished and partitioned into C and D, where C is a set of condition attributes and D is decision attribute, such that S=(U,C∪D). For every a∈C, a:V→Va, where Va represents a set of values of the attribute a, which called the domain of attribute a.

Definition 3.2: let S=(U,C∪D) for any B, where B is a subset of C, there is an equivalence relation, which represents a binary relation called indiscernibility relation and defined as

IND(B)={(x, y)∈U*U:a(x)=a(y) for all a in B} (1)

Definition 3.3: let S=(U,C∪D) , B⊆C and X⊆U. The lower approximation set $\underline{B}(X)$ is a set of all instances in U that can be surely classified as elements that belongs to X when using B

$$\underline{B}(X) = \{x \in U | [x]_B \subseteq X\} \quad (2)$$

And the upper approximation set $\overline{B}(X)$ is a set of elements that not surely classified as elements that belongs to X when using B, and can be defined as

$$\overline{B}(X) = \{x \in U | [x]_B \cap X \neq \emptyset\} \quad (3)$$

The boundary region of X expresses the uncertainty of knowledge, and defined as

$$BN_B(X) = \overline{B}(X) - \underline{B}(X) \quad (4)$$

The accuracy of approximations is defined as

$\alpha_B(X) = \left|\underline{B}(X)\right| / \left|\overline{B}(X)\right|$ (5)

$0 \leq \alpha_B(X) \leq 1$, if $\alpha_B(X) = 1$ then X is crisp otherwise X is rough.

Definition 3.4: let S=(U,C∪D), x∈U, the membership of x to X given B is defined as

$\mu_x^B(x) = \frac{|X \cap B(x)|}{|B(x)|}$ (6)

Where $\mu_x^B: U \rightarrow < 0,1 >$, and $|X|$ symbolizes the cardinality of X.

Definition 3.5: let S=(U,C∪D), the dependency between C and D, symbolized C⇒D, defined as

$k = \gamma(C, D) = \frac{POS_C(D)}{|U|}$ (7)

Where 0≤k≤1 and $POS_C(D) = \cup_{X \in U/D} \underline{C}\,(X)$. If k=1 this means that D depends totally on C, otherwise D depends partially on C.

Definition 3.6: let S=(U,C∪D), and a∈ C , a is dispensable in C if

$\gamma(C, D) = \gamma(C - \{a\}, D)$ otherwise a is indispensable.

If all the attributes in C are indispensable, then C will be called independent.

Let $\acute{C}$ be a subset, where $\acute{C} \subseteq C$, $\acute{C}$ is reduct of C if

$\gamma(C, D) = \gamma\left(\acute{C}, D\right)$ (8)

Definition 3.7: let S=(U,C∪D), and a is an attribute, where a∈ C. the significance of attribute a is defined as

$\sigma_{(C,D)}(a) = \frac{(\gamma(C,D) - \gamma(C-\{a\},D))}{\gamma(C,D)} = 1 - \frac{\gamma(C-\{a\},D)}{\gamma(C,D)}$ (9)

Where $0 \leq \sigma(a) \leq 1$. Let B a subset of attributes, B⊆C, the significance of B is defined as

$\sigma_{(C,D)}(B) = \frac{(\gamma(C,D) - \gamma(C-B,D))}{\gamma(C,D)} = 1 - \frac{\gamma(C-B,D)}{\gamma(C,D)}$ (10)

If B is reduct of C, then $\sigma_{(C,D)}(B)=0$, Any subset of C is called the approximate reduct, and any set has a number, called error of reduct approximation and can be defined as

$$\varepsilon_{(C,D)}(B) = \frac{\gamma(C,D) - \gamma(B,D)}{\gamma(C,D)} = 1 - \frac{\gamma(B,D)}{\gamma(C,D)} \quad (11)$$

Symbolized as $\varepsilon(B)$, and demonstrates how the subset B approximates the set C (condition attributes). Clearly $\varepsilon(B) = 1 - \varepsilon(C - B)$ and also $\varepsilon(B) = 1 - \sigma(B)$.

4. The proposed feature selection algorithm

The proposed feature selection algorithm can be explained by three subsections, which includes the proposed measure based on modified rough membership function that will be used to calculate the information amount by depending on the modified attribute membership, selection method and stopping condition.

4.1 The proposed measure based on modified rough membership function

The proposed feature selection algorithm offers a new measure based on rough membership function, measures the information amount of each value in an attribute by calculating its membership with the decision attribute, where the information

amount of an attribute value can also consider as a describer of the value effectiveness.

The new measure starts by measuring the effectiveness or the information amount of each value in an attribute by calculating the sum of the value membership with every class in the decision attribute, but rough membership function is modified to be suitable for differentiating between the informative value and the other values of an attribute, and consequently can recognize the useful or important attribute from the others. The conventional membership function cannot be used in this measure, because the membership of any value with all the decision attribute classes is equal to one and this result will never be changed from one value to another. So, the value of the whole attribute membership will be equal to the number of attribute values.

For example, let consider there is a decision table (table (1)) consists of two condition attributes (A1, A2) and one decision attribute (D), each condition attribute has some different values, also the decision attributes has many classes, these attributes with its values are presented in the table as follows:

Table (1): a simple decision table

	A1	A2	D
x1	1	3	Y
x2	1	1	N
x3	2	1	N
x4	1	2	N
x5	2	2	Y

So, to calculate the membership of the whole attribute A1, first calculate the membership of each value in A1 with every decision attribute classes by using rough membership function as follows:

For value=1 with the class=Y , $\mu_x^{A1_1}(x) = 1/3$

For value=1 with the class=N , $\mu_x^{A1_1}(x) = 2/3$

From above, the membership of the value =1 with the decision classes can be obtained by finding the summation of its membership with each decision class, by using the formula:

$$V_i M = \sum_{j=1}^{m} \frac{|ai(x) \cap X|}{|ai(x)|} \ldots . (12)$$

Where V_iM is the attribute value (ai) membership and m is the number of decision classes, so $V_1M = \frac{1}{3} + \frac{2}{3} \rightarrow V_1M = 1$,

And by using the same way of calculation, the membership of the value=2 can be obtained, which will also be equal to1. In order to use the membership of attribute values as a measure to express the information amount or the effectiveness of the attribute, the following formula will be used for this purpose:

$$AM_a = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|ai(x) \cap X|}{|ai(x)|} = \sum_{i=1}^{n} V_iM \ldots . (13)$$

Where $AM_a$is the information amount of the attribute (a) based on membership function and n is the number of attribute values. To measure the effectiveness of attribute A1, the previous formula will be applied on A1 as follows:

$$AM_{A1} = \sum_{i=1}^{2} V_iM$$

$$= 1 + 1$$

$$= 2$$

From above, one can observe that the information amount of attribute A1 is equal to the number of the attribute values, and that means this formula cannot be used to express the importance of attributes because it cannot express the effectiveness of attributes or the strength of the dependancy between the condition attribute and the decision attribute, there for the membership function is modified to be suitable for measuring the effectiveness of each value within an attribute.

The modified rough membership function (denoted $M\mu_x^B(x)$) will be as the following formula:

$$M\mu_x^B(x) = \frac{|(B(x) \cap X)^2|}{|B(x)|} \ldots . (14)$$

Where B represents one or more of condition attributes. This formula can be used to express the membership of attribute value with all decision classes (denoted MV_iM) as follows:

$$MV_iM = \sum_{j=1}^{m} \frac{|(ai(x) \cap X)^2|}{|ai(x)|} \ldots . (15)$$

$MV_iM\ value \in [1, z]$, where z is the number of elements that have the same attribute value (number of elements in elementary set of the attribute value).

This formula aims to increase the disparity between a value and another within an attribute according to its certainty. If the attribute value has a magnitude of

certainty with a specified decision class, then its membership value will be maximized. When a value of an attribute has a high level of certainty with a decision class, the maximization of its membership value will be increased and vice versa. To explain this aim, the formula will be applied on the attribute A1 from the table (1).

The membership of the value =1 with class=Y, $M\mu_x^{A1_1} = \frac{1^2}{3} \rightarrow M\mu_x^{A1_1} = \frac{1}{3}$

The membership of the value =1 with class=N, $M\mu_x^{A1_1} = \frac{2^2}{3} \rightarrow M\mu_x^{A1_1} = \frac{4}{3}$

So, $MV_1M = \frac{1}{3} + \frac{4}{3} \rightarrow MV_1M = \frac{5}{3} \cong 1.66$

The membership of the value=2 was also obtained by using the same way of calculation, and it was equal to 1 ($MV_2M = 1.0$).

From the above results, one can observe that the value=1 has a membership value higher than the membership of the value =2 because the value=1 include elements have certainty higher than the certainty of the value=2 elements.

As mentioned before, the modification of the membership function is required in order to measure the effectiveness or the information amount of the attribute, therefore the previous formula will be used to express the information amount of the whole attribute as follows:

$$MAM_a = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|(ai(x) \cap X)^2|}{|ai(x)|} = \sum_{i=1}^{n} MV_i m \ldots . (16)$$

Where MAM_a is the modified attribute membership. The formula represents a measure of the attribute effectiveness according to the information amount obtained from the attribute membership.

The values that obtained after applying the formulated measure on the attributes in the table (1) are MAM_{A1}=2.66 and MAM_{A2}=4, which shows that the attribute A2 has a higher information amount than attribute A1, and that leads to consider the attribute A2 is more important than A1 because it has elements with a higher level of certainty than the elements certainty of the attribute A1.

Also, one can measure and find the effectiveness of a subset of attributes by using the modified membership as follows

$$MAM_P = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|(B_i(x) \cap X)^2|}{|B_i(x)|} \ldots \ldots (17)$$

4.2 The selection method

These measures will represent a base of the feature selection algorithm. The feature selection algorithm first computes the effectiveness of each condition attribute by calculating its information amount; where the best attribute is a one that has a higher information amount among other attributes.

Computing the effectiveness of each attribute will be simplified. By the equation (16), one can calculate the information amount that expresses the effectiveness of each condition attribute by depending on the information amount that obtained from its values.

The next step in the algorithm is selecting best condition attribute, which gives the highest effectiveness value (highest information amount) when it's combined with other attributes that exist in the set of selected features and add it to the set of selected features, or have a higher effectiveness as a single attribute if there no features are selected yet. If there is two or more attributes have highest effectiveness when it combined with the previously selected attributes, then the algorithm will select the attribute that have highest effectiveness when it single.

The selected features set will be considered as conditions to satisfy the decision in the set of extracted rules. The algorithm continues selecting attributes and adding it to the selected features set until the stopping condition is met.

4.3 Stopping condition

As it's known, reduct represents a minimal subset of features (condition attributes) that can be used to achieve the same classification of decision table instead of using the whole set of features, where the unused features are superfluous, or unimportant to be used in the classification process. So, reduct will be used in this algorithm as a stopping condition to prevent adding the unimportant or less effective features to the set of left hand side conditional attributes of the rules. Reduct occurs when the significance of the unselected attributes is equal to zero, or in another ward, the stopping condition is met when the ratio of calculating the dependency of the unselected conditional attributes with the decision to the dependency of all conditional attributes with the same decision equal zero.

The stopping condition is met when the attributes that are not belonging to the reduct (selected attributes) has significance equal to zero, that means the unselected attributes are unimportant and can be discarded. In other words, stopping condition is met when the result that obtained from the equation (11) is equal to zero.

The main steps of the proposed feature selection method are shows in figure 1

Algorithm 1: feature selection algorithm

Input: C, condition attribute;
D, decision attribute;
Output: B, selected subset of features;

1) B={}
2) While significance of C $\neq 0$
3) Select attribute {a} from C if (B∪ {a}) has highest $MAM_{B\cup\{a\}}$ value using Eq. 17
4) If there is more than attributes have a highest $MAM_{B\cup\{a\}}$ value Then select the attribute that has a highest MAM_a value using Eq. 16
5) B=B∪ {a}
6) C=C-{a}
7) Return B

Figure(1): feature selection algorithm based on modified membership function.

5. Experimental results

In this section, the experimental results of the proposed algorithm are presented. The comparison between the original size of the data and the size of data after selecting features is introduced, and accuracy also compared. Four different data sets are used to evaluate the performance of the proposed algorithm, includes dataset of the terrorist attacks in Iraq, weather in Iraq, and two medical data, one of diabetes disease and the other of heart disease.

The first comparison in this section is between the attribute size before and after using the feature selection algorithm. The result of this comparison is given in table 2.

Table (2): the attributes size before and after selecting features.

Name of dataset	Original attribute size	Attributes size after FS
Terrorist attacks	4	4
Weather	11	7
Diabetes	7	6
Heart disease	12	9

From above one can see that the proposed feature selection algorithm succeeded in reducing the attributes size of the most used datasets except the terrorist attacks dataset, because the terrorist attacks dataset are independent.

The other comparison shows the number of the generated rules by rough set theory before and after using the feature selection algorithm, which given in table 3.

Table (3): number of the generated rules before and after selecting features.

Name of dataset	No. of rules before FS	No. of rules after FS
Terrorist attacks	1460	1460
Weather	151	135
Diabetes	41	35
Heart disease	212	205

The proposed feature selection algorithm minimizes the number of attributes, which will be used for generating rules. When only the effective attributes are used for generating rules, that will contribute in presenting a less number of rules and less complexity of rules, and also removes the redundant values which effects on the accuracy of generated rules.

The final comparison is between the accuracy of classification by using rough set before and after selecting features, which given in table 4.

Table (4): the accuracy before and after selecting features.

Name of dataset	Accuracy before FS	Accuracy after FS
Terrorist attacks	71.71%	71.71%
Weather	14.58%	29.16%
Diabetes	70%	85%
Heart disease	76.38%	79.16%

From above, one can see that the accuracy is increased and that because of removing some redundant attributes, which was decreasing the accuracy of the generated rules. The accuracy of the rules that generated from the terrorist attacks doesn't effected, because no change in the number of attributes that used in the generated rules.

6. Conclusion

To select the effective features from the complete set of features and reducing the dimensionality of data, this paper proposed a feature selection algorithm using modified rough membership function, where this function is used for finding the effectiveness of each attribute value, and then finding the effectiveness of each attribute by using the formulated modified attribute membership. Four different datasets are used in the experiments. The results, analysis, and discussion shows that the proposed feature selection algorithm can effectively reducing the dimensionality of data and increasing the accuracy of the generated rules when it used with rough set theory and decreasing the number of the generated rules.

References

[1] R. Kohavi, G. H. John, **''Wrappers for feature subset selection''**, Artificial intelligence. Vol. 97 , pp. 273-324, 1997.

[2] J. Yang, V. Honavar, **''Feature subset selection using a genetic algorithm, In: Feature extraction, construction and selection''**, Springer US, pp. 117-136, 1998.

[3] I. A. Gheyas, L. S. Smith, **''Feature subset selection in large dimensionality domains, Pattern recognition''**, Vol. 43 ,pp. 5-13, 2010.

[4] J. Komorowski , L. Polkowski and A. Skowron , "**Rough sets: A tutorial**", Springer, 1999.

[5] Richard Jensen , "**Combining rough and fuzzy sets for feature selection**". PhD thesis, School of Informatics, University of Edinburgh, Scotland, 2005.

[6] A. Chouchoulas, Q. Shen, **''Rough set-aided keyword reduction for text Categorization''**, Applied Artificial Intelligence. Vol. 15, pp. 843-873, 2001.

[7] L. Sun, Et al.," **Rough Entropy-based Feature Selection and Its Application**", Journal of Information & Computational Science, 2011.

[8] C. Velayutham and K. Thangavel," **Rough Set Based Unsupervised Feature Selection Using Relative dependency Measures**", Journal of Computational Intelligence and Informatics, Vol. 1- No. 1, June 2011.

[9] K.Anitha1, Dr.P.Venkatesan2،" **Feature selection by rough –quick reduct algorithm**"، International Journal of Innovative Research in Science, Engineering and Technology، Vol. 2, Issue 8, August 2013.

[10] T. Sridevi and A. Murugan, **''A Novel Feature Selection Method for Effective Breast Cancer Diagnosis and Prognosis**", International Journal of Computer Applications, Vol. 88 – No.11, February 2014.

[11] P. Mahajan, R. Kandwal and R. Vijay," **Rough Set Approach in Machine Learning: A Review**", International Journal of Computer Applications, Vol. 56– No.10, October 2012.

[12] Z. Pawlak and A. Skowron, "**Rudiments of Rough Sets**", Elsevier, Vol. 177- No. 1, pp. 3 – 27, January 2007.

[13] www.iraqbodycount.org

[14] http://mldata.org/repository/data/viewslug/datasets-uci-heart-c/