Redefining core preliminary concepts of classic Rough Set Theory for feature selection

Muhammad Summair Raza, Usman Qamar *
Department of Computer Engineering, College of Electrical & Mechanical Engineering (E&ME), National University of Sciences and Technology (NUST), Pakistan

A R T I C L E   I N F O
Keywords:
Feature selection
Rough Set Theory
Reduct
Dependency
Lower approximation
Upper approximation

A B S T R A C T
Data is growing at an exponential pace. To cope with this data explosion, we need effective data processing and analysis techniques. Feature selection is selecting a subset of features from a dataset that still provides most of the useful information. Various tools are available as underlying framework for this process however, Rough Set Theory is the most prominent tool due to its analysis friendly nature. Majority of Rough Set based feature selection algorithms use positive region based dependency measure as the sole criteria to select feature subset. Calculating positive region requires calculation of lower approximation which consequently involves indiscernibility relation. In this paper, new definitions of two Rough Set preliminaries i.e. lower and upper rough set approximation are proposed. New definitions of approximations are computationally less expensive as compared to the conventional. The proposed redefinitions showed 42.78% decrease in execution time for redefined lower approximation and 43.06% decrease in case of redefined upper approximation, for five publicly available datasets while maintaining 100% accuracy. Finally based on these redefined approximations we proposed a feature selection algorithm, which when compared with state of the art techniques showed significant increase in performance without the affecting the accuracy.

1. Introduction
In the current era, data growth is exponential as it nearly doubles every two years and by year 2020 it will be 44 trillion gigabytes i.e. ten times its size in 2014 (The Digital Universe, 0000). This is called data explosion which has resulted in a significant increase in number of samples. However, the data is useless if it cannot be processed and knowledge cannot be extracted for decision making. This leads towards the need of a mechanism to help us accurately extract useful information (Pes et al., 2017). Here feature selection is an effective method which can be applied to get a selective subset of features that can be used on behalf of the entire dataset. It is the process of selecting subset of features that still provides the most of useful information (Richard and Shen, 2008). Thus a good feature selection algorithm helps us select optimum feature subset i.e. feature subset with minimum number of features that provide maximum of the information present in the dataset.

Main purpose of feature selection is to remove unnecessary features. Unnecessary features can be classified as irrelevant features and redundant features (Richard and Shen, 2008). Irrelevant features are the ones that do not have any effect on the target concept, redundant features, on the other hand, do not add any information to the target concept, instead they may affect the performance and accuracy of models based on such data. Thus the features resulted by the feature selection algorithms have high correlation with the target concept.

However, feature selection is not any easy task. There are three ways in which feature selection algorithms can perform their job. The simplest way is to evaluate all possible subsets i.e. all combination of features to find the optimal one, however, exhaustive search is not practical because it requires large amount of resources as searching a dataset with “n” number of attributes, exhaustive search requires to evaluate $2^n$ features, which may only be possible for smaller datasets. An alternative way is to use random search where a candidate feature subset can be randomly generated (Hua et al., 2009). Algorithm performs in iterations, on each iteration and random feature set is selected and its fitness is evaluated. The process continues until we find a feature subset whose fitness fulfills the required criteria. The problem with such algorithms is that they may contain redundant attributes. Third and most commonly used method uses heuristic approach (Hua et al., 2009) where some heuristics function is used to guide the search. The best algorithm is the one that gives optimum feature subsets with highly relevant and non-redundant features without effecting the performance of the algorithm.
Various feature selection techniques (Varma et al., 2016; Zhao et al., 2016; Koppers et al., 2016; Zhou et al., 2016; Bertolazzi et al., 2016; Xuan et al., 2017; Nivetithia et al., 2016; Huang, 2014; Majumdar et al., 2016) have been proposed in literature in effort to provide efficient and effective feature selection algorithm.

Recently Rough set theory (RST) has always been a prominent tool for feature selection due to its analysis friendly nature. Proposed by Pawlak (1982) and Pawlak and Skowron (2007), it provides many structures and tools to store and analyze data and has been used in many domains so far e.g. image processing (Shi et al., 2016; Jothi, 2016), data mining (Prasuna et al., 2016; Yao et al., 2016), banking and finance (Bai et al., 2016) etc. Various feature selection algorithms have been proposed based on RST e.g. Das and Roy (2016), Huang et al. (2016), Dai et al. (2016) and Singh et al. (2016). However, majority of algorithms use conventional positive region based dependency measure. The conventional dependency method uses positive region to calculate dependency of attribute “D” on attribute “c”. Calculating positive region requires calculation of indiscernibility relation i.e. equivalence class structure which is computationally expensive method and thus becomes a bottleneck for performance of feature selection algorithms.

In this paper we have proposed a new feature selection algorithm by avoiding the calculation of computationally expensive indiscernibility relation based dependency measure. Research was performed in three steps, in first step new definitions of two Rough Set preliminaries i.e. lower and upper rough set approximation were given. The new definitions are semantically equal to the conventional definitions but provide computationally efficient method to calculate these approximations. In second step new dependency calculation method was proposed based on the redefined preliminaries. Finally, in third step, the new feature selection method was proposed based on new dependency calculation approach.

Rest of the paper is organized as, Section 2 presents related work. In Section 3 Rough Set based preliminary concepts are presented. Proposed solution is discussed in Section 4. Results are analyzed and discussed in Section 5, while the conclusion is presented in Section 6.

2. Related work

Various approaches have been proposed for feature selection using rough set theory. Almost all of them use conventional indiscernibility based lower approximation to calculate dependency. Here we have presented few in Table 1.

However, all of these approaches use positive region based conventional dependency measure which degrades the performance of algorithms and makes them computationally too expensive. In the proposed dependency based measure, we have avoided the calculation of this computationally expensive positive region calculation step.

3. Rough Set Theory

Rough Set Theory (RST) was proposed by Pawlak (Bertolazzi et al., 2016; Xuan et al., 2017). Some basic concepts of RST are discussed in this section. Further details of RST can be found in Xuan et al. (2017).

An “Information System (IS)” is a simple mechanism to represent the knowledge. It is just like a relational table with rows representing the objects, entities or records and columns representing the attributes or features.

A decision system $A = (\{U, C \cup \{D\}\})$ is an information system having decision class “D”. It is a dataset where “U” represents non empty finite set of objects (called universe) and “C” is the nonempty finite set of attributes such that $c : U \rightarrow V_c$, where $V_c$ is value set for $c \in C$. Table 2 shows a sample Decision System.

Let $A = (\{U, C \cup \{D\}\})$ be a decision system, the indiscernibility relation $IND_A(C)$ also denoted by $\{x\} A$ defines the set of objects that are indiscernible or indistinguishable w.r.t information contained in $C$.

$$IND_A(C) = \{(X_1, X_2) \in U^2 | \forall c \in C \{X_1\} = \{X_2\}\}. \quad (1)$$

So:

$$IND_A(\{\text{Experience}\}) = \{(\{X_1, X_4, X_5\}, \{X_2\}, \{X_3\}\})$$

$$IND_A(\{\text{Qualification}\}) = \{(\{X_1, X_5, X_3\}, \{X_2, X_4\}\}).$$

Let $X \subseteq U$ be any arbitrary set of objects. Normally it is not possible to define such a set in crisp manner. RST provides two important concepts called lower approximation and upper approximation to define such sets.

For $X \subseteq U$ and $B \subseteq C$. B-lower approximation $BX$ and B-upper approximation $\overline{BX}$ can be defined as:

$$BX = \{X | X \subseteq B \subseteq X\} \quad (2)$$

$$\overline{BX} = \{X | X \cap B \neq 0\} \quad (3)$$

B-Lower approximation, which is also called positive region, defines the objects that with respect to information contained in $B$, will surely belong to $X$. On the other hand $B$-Upper approximation defines objects that with respect to information contained in $B$ can possibly belong to $X$.

Boundary Region $BN_B(X) = \overline{BX} - BX$ contains the objects that we cannot definitely classify into $X$ with respect to $B$.

For example we consider the concept, $X = \{x | \text{job} (x) = \text{yes}\}$. The objects belonging to lower approximation, by using attributes $B = \{\text{Experience}, \text{Qualification}\}$ will be:

$$BX = \{X3, X4\}.$$

Similarly objects belonging to upper approximation will be:

$$\overline{BX} = \{X1, X3, X4, X5\}.$$

Finally the objects belonging to boundary region will be:

$$\overline{BX} - BX = \{X1, X5\}.$$

Another important measure provided by RST for the sake of data analysis is attribute dependency. The dependency of an attribute “D” on “C” defines that how uniquely “C” determines “D”. For a decision system $A = (U, C \cup \{D\})$, the attribute “D” depends on “C” by the degree “K”, calculated as:

$$k = \gamma(C, D) = \frac{POS_{S}(D)}{|U|} \quad (4)$$

where

$$POS_{S}(D) = \bigcup_{x \in U/D} C(X) \quad (5)$$

$POS_{S}(D)$ is called positive region of “U/D” w.r.t. “C”, if $k = 0$, “D” does not depend on “C”, for $0 < k < 1$, “D” depends on “C” partially where as if $K = 1$, $D$ fully depends on $C$.

For $R \subseteq C$, a is Reduct if:

$$\gamma(R, D) = \gamma(C, D)$$

i.e. the dependency of decision attribute on Reduct is same as that of its dependency on entire feature set.

4. Proposed solution

In this paper, we have proposed a new feature selection algorithm based on new dependency calculation method. Here dependency calculation method is based on new definitions of Rough Set approximations. All these steps are explained with example using the dataset given in Table 3.

4.1. Redefined approximations

We have redefined both lower and upper approximations. The new definitions are semantically same to the conventional definitions but provide computationally more efficient method for calculating these approximations by avoiding equivalence class structure.
دریافت فوری
متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات