Information Sciences 343–344 (2016) 41–65

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An incremental dependency calculation technique for feature selection using rough sets

Muhammad Summair Raza, Usman Qamar*

Department of Computer Engineering, College of Electrical & Mechanical Engineering (E&ME), National University of Sciences and Technology (NUST), Pakistan

ARTICLE INFO

Article history:
Received 8 January 2015
Revised 21 October 2015
Accepted 21 January 2016
Available online 30 January 2016

Keywords:
Rough set theory
Feature selection
Dependency
Reducts
Genetic algorithms

ABSTRACT

In many fields, such as data mining, machine learning and pattern recognition, datasets containing large numbers of features are often involved. In such cases, feature selection is necessary. Feature selection is the process of selecting a feature subset on behalf of the entire dataset for further processing. Recently, rough set-based approaches, which use attribute dependency to carry out feature selection, have been prominent. However, this dependency measure requires the calculation of the positive region, which is a computationally expensive task. In this paper, we have proposed a new concept called the "Incremental Dependency Class" (IDC), which calculates the attribute dependency without using the positive region. IDCs define the change in attribute dependency as we move from one record to another. IDCs, by avoiding the positive region, can be an ideal replacement for the conventional dependency measure in feature selection algorithms, especially for large datasets. Experiments conducted using various publically available datasets from the UCI repository have shown that calculating dependency using IDCs reduces the execution time by 54%, while in the case of feature selection algorithms using IDCs, the execution time was reduced by almost 66%. Overall, a 68% decrease in required runtime memory was also found.

1. Introduction

The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features [10]. In real-world problems, feature selection is a must due to the abundance of noisy, irrelevant or misleading features [3,11,12]. Various techniques have been proposed in the literature as attempts to provide a suitable feature selection method. These include feature selection using neural-fuzzy classifiers [1], instance-based learning [3], Principal Component Analysis [7], Bayesian feature selection [5], Ant Colony Optimization-based methods [15,26], feature selection in multi-dimensional time series [8], correlation- and instance-based feature selection [16], heuristic methods for finding critical dimensions [25] and boundary region-based feature selection algorithms [17].

Rough set theory (RST), proposed by Pawlak [[20]] and Pawlak and Skowron [[21]], has been used by many researchers as the underlying framework for feature selection. It is a mathematical tool to handle imperfection in knowledge, i.e., imprecision, uncertainty and vagueness. It has been applied in many domains, such as economy and finance [22], medical diagnosis [9], medical imaging [29], interactive computation [12] and data mining [18]. Different algorithms have been proposed on

* Corresponding author. Tel.: +92 3235577710.
E-mail addresses: summair.raza@ceme.nust.edu.pk (M.S. Raza), usmanq@ceme.nust.edu.pk (U. Qamar).

http://dx.doi.org/10.1016/j.ins.2016.01.044
0020-0255/

the basis of the concepts provided by rough set theory. Set approximation and dependency calculation are basic steps towards finding the relevant features (reducts) from the original dataset while still maintaining relevant information. Rough set-based dependency measures require calculating the positive region, which is a computationally expensive solution to the problem and is only practical for simple datasets (the positive region will be discussed in Section 2).

1.1. Positive region-based approaches

In [10], Inbarani et al. presented a supervised hybrid feature selection algorithm based on particle swarm optimization (PSO) and rough sets. An initial population of particles is constructed with random positions and velocity. The fitness function of each particle is evaluated using a conventional dependency measure. The algorithm then selects an attribute with a higher dependency value and computes the fitness of the selected feature with different combinations. If the particle's fitness is higher than the previous best value within the current swarm (pbest), then this particle becomes the current best. Then, its fitness is compared with the population's overall previous best fitness (gbest). If the particle's fitness is better than gbest, gbest is updated with this position. This position represents the best feature subset obtained thus far. Finally, the velocity and position of the particle are updated. The algorithm uses a positive region-based dependency measure to calculate the dependency of the decision attribute on the conditional attributes, which is suitable only for smaller datasets and becomes a performance bottleneck for larger ones.

In [31], the authors present a rough set-based genetic algorithm (GA). The output of the algorithm was provided to an artificial neural network classifier for further analysis. The algorithm uses a dependency measure as the fitness score of the chromosomes. A stopping criterion was defined on the basis of the same fitness value. The positive region-based dependency measure used in calculating the fitness degrades the performance for large chromosomes and populations.

In QuickReduct (QR) [13], the authors attempt to develop a forward feature selection mechanism without exhaustively generating all possible subsets. The algorithm starts with an empty set and adds the attributes that result in the greatest increase in the degree of dependency. The process continues until the maximum possible value is achieved. The algorithm calculates the dependency of the attribute and selects the best candidate. If at any stage the dependency of the attribute set becomes equal to that of the entire dataset (ideally 1), the algorithm stops. The algorithm also uses a positive region-based dependency and suffers all of the drawbacks inherent to this approach.

ReverseReduct [13] is another strategy for attribute reduction; however, in contrast with the forward selection process, ReverseReduct performs backward elimination. The process starts with a reduct set comprising all of the attributes. Attributes are removed from the set incrementally until the removal of further attributes introduces inconsistency. However, this algorithm also suffers from the same issues as discussed above.

Wenbin et al. [28] present an incremental feature selection algorithm (IFSA) for feature subset selection. It starts with an original feature subset P, computes the new dependency function in an incremental manner and then checks whether P is a required feature subset or not. If the new dependency function under P is equal to that under the entire feature set, P is also the new feature subset; otherwise, a new feature subset is computed. Features with the highest significance are gradually selected and added to the feature subset. Finally, redundant features are removed to ensure the optimal output. Again, the algorithm uses a positive region-based dependency measure, thus making it unsuitable for large datasets.

Chen et al. [4] present a rough set-based feature selection method using Fish Swarm Algorithm (FSA). The algorithm starts with an initial population (swarm) of fish searching for food. Each fish represents a candidate solution. Over time, they change their positions, communicate with each other and search for the local best and global best positions. When a fish achieves maximum fitness, it perishes after obtaining the rough set Reduct. The next iteration starts after all of the fish have perished. The algorithm halts when it obtains the same feature reducts under three consecutive iterations or reaches the maximum iteration condition. The algorithm uses the same rough set-based dependency measure and thus can suffer from the same dilemma of performance degradation for large datasets.

Inbarani et al. [11] propose a feature selection method based on QuickReduct and an improved harmony search algorithm (RS-IHS-QR). This algorithm imitates the music improvisation process in which each musician improvises their instrument's pitch by searching for a perfect state of harmony. The algorithm stops when it reaches the maximum number of iterations or finds a harmony vector with maximum fitness. It uses a rough set-based dependency measure as its objective function to measure the fitness of the harmony vector, which again is a performance bottleneck for larger datasets.

Table 1 summarizes these approaches, along with the advantages and disadvantages of each.

1.2. Alternative approaches

Many approaches have been proposed in the literature to overcome the computationally expensive task of calculating the positive region.

In [14], Jiang and Yu proposed a compact discernibility information tree (CDI-tree) for attribute reduction. The CDI-tree has the ability to map non-empty elements into one path and allow numerous non-empty elements to share the same prefix, which is recognized as a compact structure to store non-empty elements in a discernibility matrix. On the base of a CDI-tree, a heuristic algorithm is also proposed. The approximate strategy of the algorithm is to delete the most unimportant attribute in each iteration. The task is performed to ensure that the algorithm can retain important attributes. At the same