Applied Soft Computing 36 (2015) 552–569

Contents lists available at ScienceDirect

Applied Soft Computing

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

Privacy preserving sub-feature selection in distributed data mining

Hemanta Kumar Bhuyan [a,*], Narendra Kumar Kamila [b]

[a] Department of Computer Science and Engineering, Mahavir Institute of Engineering and Technology, Bhubaneswar, India
[b] Department of Computer Science and Engineering, C.V. Raman College of Engineering, Bhubaneswar, India

ARTICLE INFO

Article history:
Received 29 November 2013
Received in revised form 12 June 2015
Accepted 16 June 2015
Available online 29 July 2015

Keywords:
Privacy
Distributed data mining
Feature selection
Fuzzy probability
Fuzzy random variable

ABSTRACT

This paper addresses the selection of sub-feature from each feature using fuzzy methodologies maintaining the privacy during collection of data from participating parties in distributed environment. Based on fuzzy random variables conditional expectation is used in which two fuzzy sets are generated using Borel set that helps to determine sub-feature within certain interval. The privacy and selection of sub-feature leading to a distinguished class is the main objective of this research work. These two problems are directly related to data mining problems of classification and characterization of feature. In many cases traditional techniques are not suitable for complex databases. However our methodology provides better way for selection of sub-features under different situations. The proposed model and techniques both presents extensive theoretical analysis and experimental results. The experiments show the effectiveness and performance based on real world data set.

1. Introduction

The feature selection has been issued in several application areas such as several active research areas, industries, hospitals, etc. for individual requirements. Many authors have proposed several data mining techniques to select best features. Some of the techniques are feature wrappers and filters method for novel feature selection [1], fuzzy support vector machines (SVM) for feature selection [2], fuzzy rough sets for attribute selection [3], high performing feature selection for text classification [4], fuzzy feature selection [5], mutual information for selecting features in supervised neural net learning [6]. Traditional method is somehow made for feature selection to find the importance of the feature for own task. But for new classification of data, the role of feature selection bears less importance than sub-feature selection. Since the sub-feature is the part of feature data, sometimes, it is difficult to create specific class or new class which is identified in this paper. The detailed concept of sub-feature is derived in next part of this section. Specifically in biological data, the sub-feature data is always uncertain in each feature or in each database for new class or specific class. For example, a person is suffering from fever. The type of fever can be detected by common symptoms based on the perception of physicians. But in many cases it fails to detect the exactness and patient may continue with such fever. Under this circumstances physician suggests to carry out other blood tests. Entire data set of tests may not help to indentify the disease. In such case if the physician minutely observes the sub features of the data set can help to classify the disease. So our focus is to sort out the sub feature that leads to new class/specific class. In this situation we have considered three categories of sub-feature data based on frequency. For example sub features are of less frequent, medium frequent and very large frequent data as shown in Table 1.

It is evident that all sub-feature data may or may not be present in particular class and it is very natural in case of biological data. The particular sub-feature data may not be involved in particular categories of frequency of data. So randomness occurs by considering the non-existence of sub-feature data in particular class for which there is still uncertainty of data in precise meaning of frequency. Thus the uncertainty of data is always characterized by fuzziness and the category of data is represented by fuzzy set. Data randomness is caused by some perception of uncertainties whereas fuzziness is brought about by dimness of perception. Since the data randomness is taken into consideration on the basis of dimness of perception, the random variable measures the data frequency under fuzziness of data which is determined by fuzzy random variable. The fuzzy random variables are random variables whose values are fuzzy number defined vaguely with degree of acceptability which are handled by rules of fuzzy logic. It is assumed that the three categories of data frequency (LF, MF, VLF) are characterized by fuzzy numbers with membership functions as in Table 1.

Different fuzzy methods have been used for classification, regression, feature selection and data mining model which are

* Corresponding author. Tel.: +91 9937935207.
E-mail address: mthmb.bhuyan@gmail.com (H.K. Bhuyan).

http://dx.doi.org/10.1016/j.asoc.2015.06.060
1568-4946/

Table 1
Category of sub-features along with corresponding fuzzy membership functional values.

Categories of sub features	Fuzzy membership functional values
Less frequent (LF)	0.1
Medium frequent (MF)	0.7
Very large frequent (VLF)	0.2

applied on several databases by different researchers. But there is very few awareness about privacy preserving sub-feature selection using fuzzy model. Thus the new idea of considering sub-feature selection with privacy based on fuzzy random variables comes into picture in this paper. The concept of sub-feature data is described as follows for better understanding.

We consider each feature values as sub-feature (f_i), if frequency of feature value is more than zero, i.e., $|f_j| > 0$. Each unique feature value is also recognized as sub-feature. Thus a feature may have many sub-features and it varies from one feature to another. To clarify the concept of sub-feature data, we consider an example as follows. Let two features $F_1 = \{7, 20, 43, 24, 20, 22, 20, 24\}$ and $F_2 = \{0.3, 1.4, 1.8, 1.4, 0.5, 2.2, 1.1, 0.5\}$ with feature values are considered for feature 'Age' and 'Bilirubin' of Hepatitis data set from UCI machine learning repository. The sub-features are recognized as $\{f_1(7, 1), f_2(22, 1), f_3(43, 1), f_4(20, 3), f_5(24, 2)\}$ for feature 'Age' and $\{f_1(0.3, 1), f_2(1.1, 1), f_3(1.8, 1), f_4(2.2, 1), f_5(0.5, 2), f_6(1.4, 2)\}$ for feature 'Bilirubin'. The first and second arguments of each set are taken as feature values and its frequency, respectively, where each feature values are sub-feature data like $\{7, 22, 43, 24, 20\}$ from F_1 and $\{0.3, 1.1, 1.8, 2.2, 1.4, 0.5\}$ from F_2, respectively.

To solve the data mining task data are collected from different parties using fuzzy model. And data privacy is maintained at both data miner and data provider level. Fuzzy probabilities are applied on collecting data from each party to securely unify the perturbation with satisfactory data privacy. Some important fuzzy factors are impacted on the quality of sub-feature selection data model like frequency of sub-feature data, fuzziness for privacy of data etc. These factors help to developing the model as well as algorithm for sub-feature selection with privacy. The analytical and experimental results of fuzzy model show the effectiveness of sub-feature selection with privacy within expected intervals.

In many databases few data may not be present, even if it is present; it is beyond certain range with respect to our expectation for which when we consider the database for computation purpose to extract the knowledge, it fails to provide the exact solution statistically. UCI machine learning repository has such type of problem in some databases and we have been using such type of incomplete database for our purpose. Our contribution is to collect all data in a network without any ambiguity from different parties. The algorithm has been developed in such a way that all parties would be forced to provide complete information. Data are being processed to generate different sub features. But all sub features must have different data values and no common data present in all sub features. Practically certain cases are there in which single data leads another class out of other data values in same sub feature. Since database contains huge data, it is difficult manually to find such data values. So we need to extract such data value statistically for solving our purpose.

1.1. Related work

Data mining application under distributed system have been focused in several areas such as peer-to-peer network systems, distributed data mining, privacy preservation in distributed system, information sharing, etc. where the participating parties get exact solution based on combined database [7]. For several computational experiments, the different database are considered to have single solution based on the data collection from different parties using different techniques such as standard algorithm for decision tree [8] in peer-to-peer systems, identification of TOP-l inner products elements in P2P network [9], collective mining of Bayesian networks [10], mining criminal networks [11], incentive compatible privacy preserving of data in distributed classification [12], etc. The techniques of above paper indicate the importance of the role of participating parties in distributed environment where each party never wants to release their private data without protection. So many researchers develop the different privacy preservation model to protect the individual or organizational data such as secure multiparty computation [13], multiparty privacy preservation distributed data mining [14], privacy preserving data publishing [15], k-anonymity and l-diversity approach for privacy preservation in social network [16], privacy preserving SOM-based recommendations on horizontally distributed data [17], etc.

Although each feature in a database has important role for different data mining computation such as classification, clustering, etc., yet it needs to filter or to select the required sensitive feature from different database. Many researchers have developed different techniques for feature selection in data mining such as feature ranking based decision border [18], privacy preserving feature selection in P2P network [19], construction of fuzzy knowledge bases incorporating feature selection [20]. Other additional fuzzy methodologies have also been considered to enhance fuzzy model for feature selection such as higher order models for fuzzy random variables [21], upper and lower probabilities induced by fuzzy random variable [22], evolutionary boosting algorithms based on fuzzy rule classifiers [23], modeling vague data with fuzzy systems under a combination of crisp and imprecise criteria [24], fuzzy sets as basis for theory of possibility [25]. Recently a mutual information based feature selection method has been developed where the subset of relevant features are effective for classification of data. This method combines both feature-feature mutual information and feature-class mutual information to find optimal subset of relevant features [26]. None of the researchers of above cited papers have developed privacy preserving sub-feature selection based on fuzzy model. This problem has been taken care of in our work. The results of experiments yield fundamental insights into the problem.

1.2. Paper layout

The rest of the paper is organized as follows. In Section 2, we discuss the preliminaries of the proposed model. Section 3 elaborates the problem statement of proposed work. Section 4 derives the knowledge based data support and explains the collection of exact data from different parties. The different phases of sub-feature data processing are explained in Section 5. The analytical and technical aspects (including algorithms) of fuzzy model for sub-feature selection have been illustrated in Section 6. In Section 7, the experimental results and its analysis are discussed with real world dataset. Lastly Section 8 concludes the paper and open discussion for future work.

2. Preliminaries

In this section, the basic concepts of fuzzy random variables and privacy preserving distributed data mining are discussed for better understanding of the proposed model. Lot of research work has been carried out on feature selection by different authors, but few researchers have taken sub feature into consideration for classification of new class/special class [27]. The primary idea is to focus on the related issues in the scenario of multiparty to release