A FRAMEWORK FOR CLASSIFICATION OF MEDICAL DATA USING BIJECTIVE SOFT SET

Salman Ahmed .G*
Research Scholar M. Tech (SE), VIT University, Vellore.
Email: salman444g@gmail.com

Received on 06-08-2016
Accepted on 27-08-2016

Abstract

Classification is one of the main issues in Data Mining Research fields. The classification difficulties in medical area frequently classify medical dataset based on the result of medical diagnosis or description of medical treatment by the medical specialist. The extensive amounts of information and data warehouse in medical databases need the development of specialized tools for storing, retrieving, investigation, and effectiveness usage of stored knowledge and data. Intelligent methods such as neural networks, fuzzy sets, decision trees, and expert systems are, slowly but steadily, applied in the medical fields. Recently, Bijective soft set theory has been proposed as a new intelligent technique for the discovery of data dependencies, data reduction, classification and rule generation from databases. In this paper, we present a novel approach based on Bijective soft sets for the generation of classification rules from the data set. Investigational results from applying the Bijective soft set analysis to the set of data samples are given and evaluated. In addition, the generated rules are also compared to the well-known decision tree classifier algorithm and Naïve bayes. The learning illustrates that the theory of Bijective soft set seems to be a valuable tool for inductive learning and provides a valuable support for building expert systems.

Keywords: Bijective soft set theory; Decision rules generation; Rules classification; soft set; Bijective soft set based classification.

1. Introduction

Difficulties of showing uncertain data in economics, engineering, environmental science, sociology, medical science, and many other fields are very important for solving practical problems[1]. The evolution of the mass of data and number of existing databases far exceeds the skill of humans to analyze this data, which makes both a need and an opportunity to
Salman Ahmed G* et al. / International Journal of Pharmacy & Technology

extract knowledge from databases. Medical databases have gathered large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Analysis of medical data is often concerned with treatment of incomplete knowledge, with management of unpredictable pieces of information and with handling of various levels of representation of data. To handle incomplete knowledge and uncertain information, Rough sets prove to be an effective tool. In this paper, we propose a novel approach for classification based on bijective soft sets[10][1] for classification of uncertain data. The rest of the paper is structured as Section 2 to 7. Section 2 describes the related work of soft set, bijective soft set theory and decision making rules using rough set. Section 3 describes the research motivation of this work. Section 4 describes the fundamental concepts of bijective soft set theory. Section 5 explains the proposed algorithm with an example. This paper converses how bijective soft set can be used for generating classification rules from a set of perceived samples of the given data set [6]. Comparative analysis of the proposed work is made with decision trees and naïve bayes classifier algorithm as given in Section 6. Finally, conclusion is given in Section 8.

2. Related Work

Soft set theory was initiated by Molodstov to solve some uncertainties among data which is not solved by traditional mathematical tools. He has shown several applications of this theory by solving many practical problems in various aspects like engineering, economics, and medical science etc [10]. Aboul Ella Hassanien proposed intelligent data analysis of breast cancer based on rough set theory[1]. In [7], Fuzzy Soft Set Based Classification is proposed for Gene Expression Data. Aktaşand Çagman have proved that every rough set may be considered as a soft set, and write rough sets as predicates[2].

Maji et al. have additional studied the theory of soft sets and used this theory for some decision making problems[8][9]. They have also introduced the concept of fuzzy soft set theory for decision making problems. Pei and Miao represented the rough set model as two soft sets [11],

Herawan and Deris devoted to revealing interconnection between rough sets and soft sets and they presented a direct proof that Pawlak’s and Iwinski’s rough sets can be considered as soft sets [5]. Bana Handaga et al. proposed numerical data classification for seven types of medical data and they provided higher accuracy [3]. K. Gong et al. have proposed the bijective soft set with its operations [4].
3. Research Motivation

The main challenges in medical data classification are its huge size and its vagueness. Many researchers are trying to achieve the higher accuracy for medical data classification by reducing or selecting the informative data among the thousands of data. For cancer classification, researchers have used some machine learning algorithms like Support Vector Machines, Principle Component Analysis etc., Rough Set Theory and Fuzzy logic. In this paper we have applied bijective soft set based for classification for breast cancer, Pima Indians diabetes and liver-disorders dataset from UCI machine learning repository.

4. Preliminaries

4.1 Soft Set Theory

In this section, we describe the basic notions of soft sets. Let U be initial universe of objects and e be a set of parameters in relation to objects in U. Parameters are often attributes, characteristics or properties of objects [10].

**Definition-1.** A pair \((f, a)\) is called a soft set over U, where \(f\) is a mapping given by

\[
f : a \rightarrow p(U).
\]

In other words, a soft set over U is a parameterized family of subsets of the universe U. For \(e \in a\), \(f(e)\) may be considered as the set of \(e\) - elements of the soft set \((f, a)\) or as the set of \(e\) - approximate elements of the soft set.

4.2 Bijective Soft set Theory

Throughout this section U refers to an initial universe, \(e\) is a set of parameters, \(p(U)\) is the power set of and \(a \subseteq e\).[4]

**Definition-2.**

Let \((f, b)\) be a soft set over a common universe U, where \(f\) is a mapping \(f : b \rightarrow p(U)\) and \(b\) is nonempty parameter set. We say that \((f, b)\) is a bijective soft set, if \((f , b)\) such that

a) \(U \subseteq b \ f(e) = U\).

b) For any two parameters \(ei, ej \in b\), \(ei \neq ej\), \(f(ei) \cap f(ej) = \emptyset\).

c) In other words, Suppose \(y \subseteq p(U)\) and \(y = \{f(e1), f(e2), \ldots, f(en)\}\), \(e1, e2, \ldots, en \in b\). From definition 2, the mapping \(f : b \rightarrow p(U)\) can be transformed to the mapping \(f : b \rightarrow y\), which is a bijective function. i.e. for every \(y \in Y\), there is exactly one parameter \(e \in b\) such that \(f(e) = y\) and no unmapped element remains in both \(b\) and \(y\).[1]
4.2.1 Operation:

**Definition-3.** (See [10]) And operation on two soft sets. If \((f, a)\) and \((g, b)\) are two soft sets then \(\langle f, a \rangle \And (g, b)\) denoted by \((f, a) \And (g, b)\) is defined by \((f, a) \And (g, b) = (h, a \times b)\), where \(H(\alpha, \beta) = f(\alpha) \cap g(\beta), \forall (\alpha, \beta) \in a \times b\).[4]

**Definition-4.** Suppose that \((f, e)\) and \((g, b)\) are two bijective soft sets over common universe \(U\). \((h, c) = (f, e) \And (g, b)\) is a bijective soft set.[4]

4.2.2 Restricted and Operation:

**Definition-5.** (Restricted And Operation on a Bijective Soft Set and a Subset of Universe). Let \(U = \{o_1, o_2, \ldots, o_n\}\) be a common universe, \(x\) be a subset of \(U\), and \((f, e)\) be a bijective soft set over \(U\). The operation of \(\langle f, e \rangle \text{ restricted AND } x\) denoted by \((f, e) \And \neg x\) is defined by \(U_e \subseteq e \{f(e): f(e) \subseteq x\}\).[4]

**Example 1.** Let \((g, b)\) be a bijective soft set over a common universe \(U, U = \{o_1, o_2, o_3, o_4\}\).

Suppose the following: \((g, b) = \{e_1 = \{o_1, o_3\}, e_2 = \{o_2, o_3\}, e_3 = \{o_4\}\}\). \(x = \{o_1, o_3\}\). From Definition 4.2.2 we can write \((g, b) \And \neg x = \{o_1, o_3\}\)

5. Bijective Soft Based on Classification

5.1 Bijective Soft set based on classification - Proposed algorithm

In the new Bijective soft set based classification algorithm, the rules generation is achieved by using And, Restricted And operations in bijective soft set theory. It starts with constructing bijective soft set and apply And, Restricted And operations the finally propose classification rules.

Algorithm Bijective – Soft set Classification (BSCLASS)

Input: Data set D

Output: A set of Rules R

Step 1: Construct bijective soft set for all conditional attributes \((f_i, e_i)\) for \(i=1\) to \(n-1\), \(n\) is the number of attributes.

Step 2: Construct bijective soft set for decision attribute \((g, b)\).

Step 3: Apply AND operation on the bijective soft set \((f_i, e_i)\). And result to be stored in \((H, C)\).

Step 4: Apply the Restricted AND operation \((h, c)\) and decision attribute \((g, b)\).

\(\langle f_i, e_i \rangle \text{ restricted AND } (g, b)\), for \(i=1\) to \(n-1\) and store the result in \((f, d)\)
5.2 Numerical Illustration

<table>
<thead>
<tr>
<th>TABLE I: SAMPLE DATA SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
</tr>
<tr>
<td>o1</td>
</tr>
<tr>
<td>o2</td>
</tr>
<tr>
<td>o3</td>
</tr>
<tr>
<td>o4</td>
</tr>
<tr>
<td>o5</td>
</tr>
<tr>
<td>o6</td>
</tr>
<tr>
<td>o7</td>
</tr>
<tr>
<td>o8</td>
</tr>
<tr>
<td>o9</td>
</tr>
<tr>
<td>o10</td>
</tr>
</tbody>
</table>

Where a1, a2, a3 represents the conditional attributes and d represent the decision attribute.

Step 1. Constructing bijective Soft set from conditional attributes

\((f_1, a_1) = \{o_1, o_4, o_5, o_6, o_9\} \{o_2, o_3, o_7, o_8, o_{10}\}\)

\((f_2, a_2) = \{o_1, o_3, o_5, o_6, o_8\} \{o_2, o_4, o_7, o_9, o_{10}\}\)

\((f_3, a_3) = \{o_1, o_3, o_4, o_6, o_8, o_9, o_{10}\} \{o_2, o_5, o_7\}\)

Step 2. Constructing bijective Soft set from Decision attributes

\(d(1) = \{o_1, o_4, o_5, o_6, o_9\}\)

\(d(2) = \{o_3, o_8\}\)

\(d(3) = \{o_2, o_7, o_{10}\}\)

Step 3. Apply And operation on the bijective soft set \((f_i, a_i)\). Table 2 shows tabular form of this And operation.

<table>
<thead>
<tr>
<th>TABLE II: TABULAR FORM OF ((f_1, a_1) \land (f_2, a_2) \land (f_3, a_3))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects</td>
</tr>
<tr>
<td>o1</td>
</tr>
<tr>
<td>o2</td>
</tr>
<tr>
<td>o3</td>
</tr>
<tr>
<td>o4</td>
</tr>
<tr>
<td>o5</td>
</tr>
<tr>
<td>o6</td>
</tr>
</tbody>
</table>
Where \( e_1 = \{o_1, o_4, o_5, o_6, o_9\} \land \{o_1, o_3, o_5, o_6, o_8\} = \{o_1, o_5, o_6\} \)

\[
\begin{array}{cccccccc}
    & o_7 & 0 & 0 & 0 & 1 & 0 & 1 \\
\hline
    o_8 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
    o_9 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
    o_{10} & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
\end{array}
\]

\( e_2 = \{o_1, o_4, o_5, o_6, o_9\} \land \{o_2, o_4, o_7, o_9, o_{10}\} = \{o_4, o_9\} \)

\( e_3 = \{o_1, o_4, o_5, o_6, o_9\} \land \{o_1, o_3, o_4, o_6, o_8, o_9, o_{10}\} = \{o_1, o_4, o_6, o_9\} \)

\( e_4 = \{o_1, o_4, o_5, o_6, o_9\} \land \{o_2, o_5, o_7\} = \{o_5\} \)

\( e_5 = \{o_2, o_3, o_7, o_8, o_{10}\} \land \{o_1, o_3, o_5, o_6, o_8\} = \{o_3, o_8\} \)

\( e_6 = \{o_2, o_3, o_7, o_8, o_{10}\} \land \{o_2, o_4, o_7, o_9, o_{10}\} = \{o_2, o_7, o_{10}\} \)

\( e_7 = \{o_2, o_3, o_7, o_8, o_{10}\} \land \{o_1, o_3, o_4, o_6, o_8, o_9, o_{10}\} = \{o_3, o_8, o_{10}\} \)

\( e_8 = \{o_2, o_3, o_7, o_8, o_{10}\} \land \{o_2, o_5, o_7\} = \{o_2, o_7\} \)

\((f_1, a_1) \land (f_2, a_2) \land (f_3, a_3) = (h, c)\)

\((h, c) = \{\{o_1, o_5, o_6\}, \{o_4, o_9\}, \{o_1, o_4, o_6, o_9\}, \{o_5\}, \{o_3, o_8\}, \{o_2, o_7, o_{10}\}, \{o_3, o_8, o_{10}\}, \{o_2, o_7\}\} \)

Step 4. Apply the restricted And operation \((h, c)\) and decision attribute \((g, d)\).

\[(h, c) \land \neg d(1) = \{\{o_1, o_5, o_6\}, \{o_4, o_9\}\}\]

\[(h, c) \land \neg d(2) = \{o_3, o_8\}\]

\[(h, c) \land \neg d(3) = \{o_2, o_7, o_{10}\}\]

\((f, b) = \{\{o_1, o_5, o_6\}, \{o_4, o_9\}, \{o_3, o_8\}, \{o_2, o_7, o_{10}\}\}\)

Step 5. Generating decision rules based on restricted And result \((f, b)\).

The following deterministic decision rules are extracted:

If \(a_1 = 1\) and \(a_2 = 1\imp d = 1\)

If \(a_1 = 1\) and \(a_2 = 2\imp d = 1\)

If \(a_1 = 2\) and \(a_2 = 1\imp d = 2\)

If \(a_1 = 2\) and \(a_2 = 2\imp d = 3\)
6. Bijective Soft Set based on Classification

Classification is a data mining technique used to predict group membership for data instances. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. In this work, classification accuracy of the proposed approach is compared to two different classifiers Naïve Byes and decision table used for computing the Precision, Recall and F-measure.

6.1 Data description

Data sets taken for experiment is obtained from UCI repository (www.ics.uci.edu/~mlearn/). The medical data sets like breast cancer dataset, Liver disorder data set and Pima Indian diabetes were taken for experimental analysis of the bijective soft set based classification approach.

7. Validation measures

7.1 Precision: Precision is a measure of the accuracy provided that a specific class has been predicted [7]. It is defined by

\[ \text{Precision} = \frac{tp}{tp+fp} \]

Where tp and fp are the numbers of true positive and false positive respectively.

7.2 Recall

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is commonly also called sensitivity, and corresponds to the true positive rate[7].

\[ \text{Recall (or) Sensitivity} = \frac{tp}{tp + fn} \]

7.3 F-measure

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score of

\[ \text{F-Measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision} + \text{Recall}} \]

Table 3, 4 and 5 depicts the performance analysis of the proposed classification algorithm on breast cancer, Pima Indians Diabetes and liver-disorders dataset. Figures 1, 2 and 3 show the comparative analysis of classification algorithms for breast cancer Pima Indians Diabetes and liver-disorders datasets. The proposed approach performs better than other two approaches.
TABLE III. PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS ON BREAST CANCER DATASET

<table>
<thead>
<tr>
<th>Accuracy Measures</th>
<th>Classification algorithm</th>
<th>Decision table</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.98</td>
<td>0.44</td>
<td>0.90</td>
</tr>
<tr>
<td>Recall</td>
<td>1</td>
<td>0.667</td>
<td>0.889</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.99</td>
<td>0.533</td>
<td>0.882</td>
</tr>
</tbody>
</table>

Figure 1. Comparative analysis of classification Algorithms for Breast cancer dataset

TABLE IV. PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS ON PIMA INDIANS DIABETES DATASET

<table>
<thead>
<tr>
<th>Accuracy Measures</th>
<th>Classification algorithm</th>
<th>Decision table</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.96</td>
<td>0.78</td>
<td>0.84</td>
</tr>
<tr>
<td>Recall</td>
<td>0.92</td>
<td>0.64</td>
<td>0.73</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.94</td>
<td>0.71</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Figure 2. Comparative analysis of classification Algorithms for Pima Indians Diabetes dataset

TABLE V. PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS ON LIVER-DISORDERS DATASET

<table>
<thead>
<tr>
<th>Accuracy Measures</th>
<th>Classification algorithm</th>
<th>Decision table</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.80</td>
<td>0.78</td>
<td>0.514</td>
</tr>
<tr>
<td>Recall</td>
<td>0.97</td>
<td>0.48</td>
<td>0.429</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.89</td>
<td>0.54</td>
<td>0.471</td>
</tr>
</tbody>
</table>
The sample rules generated from liver-disorders data set is depicted in Table 5.

Attribute information:

1. a1 - mean corpuscular volume
2. a2 - alkaline phosphatase
3. a3 - alanine amino transferase
4. a4 - aspartate amino transferase
5. a5 - gamma-glutamyl transpeptidase
6. a6 - number of half-pint equivalents of alcoholic beverages drunk per day
7. d - selector field used to split data into two sets 1, 2.

The following deterministic decision rules are extracted:

If a1=78 and a2=69 and a3=24 and a4=18 and a5=31 and a6=0.5 => d=1.
If a1=99 and a2=86 and a3=58 and a4=42 and a5=203 and a6=6 => d=1.
If a1=65 and a2=63 and a3=19 and a4=20 and a5=14 and a6=0.5 => d=2.
If a1=82 and a2=55 and a3=18 and a4=23 and a5=44 and a6=8 => d=2.
If a1=85 and a2=77 and a3=45 and a4=36 and a5=77 and a6=0.8 => d=1.

Table VI SAMPLE RULES GENERATED FROM LIVER-DISORDERED DATA SET

<table>
<thead>
<tr>
<th>rule no.</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>a4</th>
<th>a5</th>
<th>a6</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78</td>
<td>69</td>
<td>24</td>
<td>18</td>
<td>31</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>99</td>
<td>86</td>
<td>58</td>
<td>42</td>
<td>203</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>85</td>
<td>77</td>
<td>45</td>
<td>36</td>
<td>77</td>
<td>0.8</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>63</td>
<td>19</td>
<td>20</td>
<td>14</td>
<td>0.5</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>82</td>
<td>55</td>
<td>18</td>
<td>23</td>
<td>44</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>
8. Conclusion

Classification of medical data is an important research problem since it helps in computer aided diagnosis. In this paper, Bijective soft set approach is proposed for generating classification rules from the data set. Comparison of the proposed approach with well-known decision table and naïve bayes classifier algorithms are made using performance metrics. Bijective soft set proves to be an effective tool for classification.

References


Corresponding Author:
Salman Ahmed .G*
Email: salman444g@gmail.com