Fuzzy Boundary Attribute Transformation in Databases

Nikhat Khan¹, Dr. Fozia H. Khan² and Dr. G. S. Thakur³

¹,²,³Maulana Azad National Institute of Technology, Bhopal

ABSTRACT

A database is a collection of data. It stores information which can be useful in the future. Fuzzy Logic is a type of logic theory which uses reasoning which is not precise or pinpoint. It is used in many ways including control systems, decision making, artificial intelligence and databases. A fuzzy boundary is an area of transition between two zones. An example is the boundary between rural and urban areas. A fuzzy boundary can also be applied to a fuzzy set, such as, the boundary between two different membership functions. In this paper, we describe the fuzzy boundary problem in databases and its solution. If a variable falls close to the boundary between two membership functions, or falls within the “fuzzy” boundary area, its membership value will be decided by the use of another variable.

Keywords: Database, Fuzzy Boundary Attribute Transformation (FBAT), Fuzzy Logic, Fuzzy Set

1. INTRODUCTION

DATABASES are a way to store data which can be useful in the future. They are used everywhere, from transport systems to banks, from local supermarkets to astronomical organizations. There are many ways in which data can be useful to us. For example data about a reservation may be required by the staff to successfully check-in a traveler. Similarly data may be stored about a customer’s preferences and it may be used again during his second visit.

Fuzzy Logic is a type of logic theory which is contradictory to binary logic [1]. It can consist of variables in between 1 and 0, which is not possible in Boolean logic. It can approximate rather than being fixed. It finds applications in many fields including electronics, cars and microprocessors. It also finds use in artificial intelligence.

A fuzzy boundary is a boundary which is not precisely marked out or is a region of transformation or transition between two distinct regions [2]. The boundary between two geographic zones is an example of a fuzzy boundary. Another example is the transition from urban to rural areas.

A fuzzy set is different from a conventional set from the fact that it can contain members that are not completely a part of the set, but are not totally non members of the set. This means that their inclusiveness in the set is partial. A conventional set has only two options for membership, true or false, but a fuzzy set also has options in between them [3].

This paper tries to decide that whether or not an object is a member of the set by the use of a second variable. This technique is named Fuzzy Boundary Attribute Transformation (FBAT). For example, a task has to done to classify the bank account balance of individuals, that is a numeric value will be given, and it has to be classified into very high, high, average, low and very low. The boundaries between the memberships have been given, and the values that fall near the boundary will be decided by the use of another value. A record is given as 395 USD. The boundary between low and average is 400 USD. This makes it quite clear that it is quite near the boundary between the two. The database also contains a value loan. It is decided that if the individual has taken a loan, he will be classified as having a low balance, in the other case, with an average balance.

For example, the statement “Temperature today is 38°C” does not explicitly state that today it’s hot, and the statement “Today’s temperature is 1.2 standard deviations about the mean temperature for daytime in the month of May” is fraught with difficulties: would a temp 1.1999999 standard deviations above the mean be hot.

The need for this technique may be felt while classifying data, or to convert numeric data into linguistic variables. It may also be used to optimize the output in case of applications like data mining, control systems.

This technique could be applied in various fields’ including customer analytics, fuzzy control systems and artificial intelligence. The advantage of this technique is that it is time saving and more accurate, as in this case, where a small difference acts as a big barrier. It takes into consideration more than one attribute. If the data derived is put to some use as marketing, where those with an average balance were to be marketed some product, this can be of use as it will optimize the list of the customers to be contacted as part of the marketing.

2. LITERATURE REVIEW

Fuzzy logic was first proposed by Zadeh L.A. of the University of California at Berkeley in 1965 in a paper [3]. He introduced the term “linguistic variables” in 1973 through his research work. The first commercial application of fuzzy logic was by a cement kiln in Denmark. Another application was made in Japan when a fuzzy control system was used to control the acceleration, braking and stopping of a train system in 1987. Subsequently, a lot of research has been done and fuzzy control systems are in wide use in electronics like washing machines, vacuum cleaners and air conditioners.
Fuzzy Logic has also been applied for decision making. Fuzzy Logic has been combined with many other techniques including soft set theory, data mining, rough sets, data warehousing and artificial intelligence [4]. Fuzzy set theory utilizes fuzzy inferencing system to reason the linguistic variables. There are number of inference systems [5] that have been developed till now. The two more prominent inference systems that are generally used are Mamdani fuzzy system [6] which is also referred as linguistic fuzzy system and the other is Takagi Sugeno fuzzy system [7]. These are mainly used for fuzzy controller.

3. AN OVERVIEW

3.1 Fuzzy Logic
Fuzzy Logic usually deals with data that is approximate which is a degree between truth and false. Fuzzy Logic was first proposed by Lotfi A. Zadeh, in 1965 [3]. Fuzzy logic has however been studied since the 1920s as infinite-valued logics notably by Łukasiewicz and Tarski [8]. Fuzzy logic is often described as always being imprecise. However this is a misconception. In large measures, fuzzy logic is accurate [9]. Boolean Logic consists of merely two possibilities, either true or false. This can be thought of as old or young. If we introduce fuzzy into this, we may get even infinitely more variables. If we introduce three, for example, we may get, young, middle-aged or old. This introduces a variable of semi-truth, or being partially old and partially young.

3.2 Fuzzy Set
Fuzzy sets are sets whose elements have a degree of membership. This means that the members can be a member or non-member of that set, as well as a partial member with a degree of membership of that set. Fuzzy sets were introduced by Lotfi A. Zadeh [3] and also by Dieter Klaua in 1965 as an extension of the classical notion of set [10]. In classical set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition — an element either belongs or does not belong to the set. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in a set; this is described with the aid of a membership function valued in the real unit interval [0, 1].

3.3 Databases
A database is an organized collection of data. The data are typically organized to model relevant aspects of reality in a way that supports processes requiring this information. Examples of database applications include computerized library systems, flight reservation systems, and computerized parts inventory systems. They store data which can be of use in the future.

3.4 The Proposed Technique
The technique proposed involves deciding the membership of an object using another value. For example, a description of an object is provided. We have to decide its membership in new, average or old. Its description provides that it is 6 years old. It is classified as 20% new and 80% average. It has also been provided that it is frequently used. Based on this, its membership in the set of average items is increased and it becomes a full member of the set of averagely old items.

In the same example, another object is given. Its description provides that it is 4 years old. It lies on the border between new and averagely old items. It is classified as 60% new and 40% averagely old. It is provided as highly used. Taking this in account, we increase its membership in the averagely old item set making it a full member of the averagely old item set.

4. METHODOLOGY
The proposed method involves the following steps-
1. Attribute Identification
2. Classification
3. Inclusion of independent attribute
4. Reclassification
5. Application

![Figure 1 The Process](image-url)

4.1. Attribute Identification
The first step involved is identification of the attributes to be classified or used. This includes two stages, selection of the dependent attributes and the selection of the independent attributes.

The dependent attributes include the ones which are to be classified. They are named as dependent as their value can be changed and is dependent on other attributes. The independent attributes are the ones which may alter the value of the...
dependent attribute. It is named independent as its value does not depend upon any other attributes. There may be more than one attribute, either dependent or independent selected.

4.2. Classification
This step includes classification of the data. This step involves deciding how much a particular object belongs to an item set, or the membership of an object in a fuzzy set. The original value provided is a numerical one. Classification is done only on the dependent attributes as only their value is set to be changed. The classes or fuzzy sets like averagely old, new or old in which the objects can be classified are also generated during this step.

4.3. Inclusion of Independent Attribute
This important step is where the independent attribute is used to classify the dependent attribute. However this happens only when the value of the attribute lies near the crisp boundary or in other words, has a partial membership between two fuzzy sets. If the membership is 100% or 1.0, then there is no point in using the independent attribute. The membership of a certain attribute is raised according to the value of the independent attribute. It may even be raised to full membership as shown in the example. The value by which membership is to be increased is also decided in this step. The rules to decide the change of value in the dependent attribute according to the independent attribute are decided first, followed by applying the rules.

4.4. Reclassification
This step involves classifying of the data again. This is not required if the membership before inclusion of the indecent attribute was 100%. This step classifies the data in normal crisp sets rather than fuzzy sets as the output should be defuzzified. The rules will remain the same as the original classification.

4.5. Application
The final step involved is the application and use of the dataset so found. The independent attributes may be removed from the new database depending on the need.

5. ADVANTAGES
The advantages of using this technique are described below-

1. Time – This technique would also save time in applications like data mining. This is due to the fact that lesser data would take lesser time in being processed.
2. Accuracy – The use of this technique has come out to be more accurate for one algorithm and otherwise has been more or less the same in other algorithms. Rather than one attribute, two or more related attributes would present more complete information about a record.
3. Resolving fuzziness – This method helps in the conversion of fuzzy sets into specific crisp sets.

6. APPLICATION
The application of the proposed technique is described in this section.

The proposed technique is applied on a database obtained from UCI Machine learning Repository and used in the paper “Using Data Mining for Bank Direct Marketing An Application of the CRISP-DM Methodology” [11] [12]. It is a bank marketing data set and contains information from a Portuguese banking institution. The proposed technique is applied to preprocess the data before it used for a data mining method, in this case classification. The data without the applied technique is also used for the same and the accuracy of the results and time spent in mining the data is compared.

The first step involved is choosing the independent and dependent attributes. The dataset chosen has a total of 17 attributes and 45211 records. An attribute balance is chosen as the dependent attribute. It has data about the customer’s balance with the bank. The independent attributes chosen are – default (customer has credit in default), housing (customer has housing) and loan (customer has unpaid loan). Balance is a continuous attribute, while the others are discrete.

The second step involved is classification. The program has been fuzzified using IF THEN rules. The results are records either wholly or partially a member of a class such as low or high balance.

The third step is inclusion of the independent attribute. This is also done using IF THEN rules. However this is done only on attributes falling under two classes.

The fourth step is reclassification. The rules are kept the same as the original classification except for that the records will be classified into crisp sets.

The fifth step is application. The data is applied to do classification (data mining). Also, independent attributes have been removed in the database on which this method has been applied as their value has already altered the dependent attribute. The program used to do this task is Waikato Environment for Knowledge Analysis (WEKA) [13]. There are various algorithms used and the results are given in Table 1. The data is also applied on data for which this technique has not been applied. The results on the same are shown in Table 2.
Table 1 : Result on Preprocessed (FBT) Data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Table</td>
<td>13.81 s</td>
<td>89.6569%</td>
</tr>
<tr>
<td>AD Tree</td>
<td>24.06 s</td>
<td>89.079%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>7.69 s</td>
<td>89.5961%</td>
</tr>
<tr>
<td>REP Tree</td>
<td>1.92 s</td>
<td>89.0873%</td>
</tr>
<tr>
<td>Random Tree</td>
<td>1.55 s</td>
<td>86.9197%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.47 s</td>
<td>87.1105%</td>
</tr>
</tbody>
</table>

Table 2 : Result on original Data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Table</td>
<td>18.03 s</td>
<td>89.6569%</td>
</tr>
<tr>
<td>AD Tree</td>
<td>28.28 s</td>
<td>89.4661%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>10.89 s</td>
<td>89.3168%</td>
</tr>
<tr>
<td>REP Tree</td>
<td>2.41 s</td>
<td>89.032%</td>
</tr>
<tr>
<td>Random Tree</td>
<td>1.67 s</td>
<td>87.2045%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.97 s</td>
<td>87.2266%</td>
</tr>
</tbody>
</table>

The accuracy for the first algorithm, decision table has been the same in both cases but the time consumed has reduced significantly. For the next one, AD Tree, the accuracy has reduced by a small value but the time taken is lesser. It is the same case with Random forest. For REP Tree, the accuracy has increased along with the time decreasing. In Random Tree, The accuracy has reduced to a small extent while the time too has reduced. For Naive Bayes, the accuracy again has reduced a bit but the time taken has reduced very significantly.

7. CONCLUSION

The proposed technique is a method to solve the boundary problem occurring in databases with precision. It utilizes fuzzy logic to improve the efficiency of the data stored in databases. This technique can find use in almost any field that utilizes databases such as data mining, artificial intelligence, data warehousing [14].

References

[3] L. A. Zadeh (1965) "Fuzzy sets". Information and Control 8 (3) 338-353

AUTHOR

Mrs. Nikhat Khan is an IT professional with over 15 years of experience. She has worked in multiple national and international organizations. Currently she is working as a Delivery Project Manager and has visited US, UK and German countries for managing and executing IT projects for renowned international clients. She is currently pursuing PhD from Maulana Azad National Institute of Technology, Bhopal (MANIT) and is eager to know about the advances in business intelligence, data mining and fuzzy logic. She has around 4 publications in National and International Journals.

Dr. Fozia H. Khan is Assistant Professor and coordinator of M. Tech (Nanotechnology) at MANIT Bhopal. Author of a book on Quantum Mechanics, Dr Fozia has guided 26 M.Tech theses and is currently guiding 7 Ph.Ds on high application materials like oxide nanomaterials for solar cells, gas sensors and other nanomaterials. With over 80 publications in national and international journals and conferences of repute, she has delivered talks in several national and international conferences, and visited King Abdul Aziz University, Jeddah, S.A, Harvard University Boston USA, and Naval Research Lab, Washington DC (USA).

Dr. G.S. Thakur has received BSc degree from Dr. Hari Singh Gour University Sagar M.P. in 2000. He has received MCA degree in 2003 from Pt. Ravi Shankar Shukal University Raipur C.G. and PhD degree from Barkatullah University, Bhopal M.P. in year 2009. He is an Assistant Professor in the department of Computer Applications, Maulana Azad National Institute of technology, Bhopal, M.P. India. He has eight year teaching and research experience. He has 26 publications in national and international journals. His research interests include Text Mining, Document clustering, Information Retrieval, Data Warehousing. He is a member of the CSI, IAENG, and IACSIT.