Fuzzy based clustering algorithm for privacy preserving data mining

Pradeep Kumar*

219, Faculty Block,
IIM Campus, Prabandh Nagar,
Off Sitapur Road, Lucknow – 226013, India
E-mail: pradeepkumar@iiml.ac.in
*Corresponding author

Kishore Indukuri Varma

Infosys Technologies Ltd.,
Survey No:210, Manikonda Village,
Rajendranagar Mandal, Lingampally,
Rangareddy District, Hyderabad 500019, India
E-mail: kishore_varma@infosys.com

Ashish Sureka

Indraprastha Institute of Information Technology (IIIT),
3rd Floor, Library Building, NSIT Campus,
Dwarka, Sector 3, New Delhi – 110078, India
E-mail: ashish@iiitd.ac.in

Abstract: Sharing of data among multiple organisations is required in many situations. The shared data may contain sensitive information about individuals which if shared may lead to privacy breach. Thus, maintaining the individual privacy is a great challenge. In order to overcome the challenges involved in data mining, when data needs to be shared, privacy preserving data mining (PPDM) has evolved as a solution. The objective of PPDM is to have the interesting knowledge mined from the data at the same time to maintain the individual privacy. This paper addresses the problem of PPDM by transforming the attributes to fuzzy attributes. Thus, the individual privacy is also maintained, as one cannot predict the exact value, at the same time, better accuracy of mining results is achieved. ID3 and Naive Bayes classification algorithms over three different datasets are used in the experiments to show the effectiveness of the approach.

Keywords: data mining; privacy preserving data mining; PPDM; fuzzy set; decision trees.


Biographical notes: Pradeep Kumar obtained his PhD from the Department of Computer and Information Sciences, University of Hyderabad, India. He also holds an MTech in Computer Science and BSc (Engg) in Computer Science.
1 Introduction

With the advances in information technology for collecting, storing and processing of data, data mining gained popularity among various organisations to meet daily business requirements. Data mining has resulted in benefits to various organisations like, medical science, financial organisations, marketing etc. Medical sciences has been benefited by data mining by achieving improvement in diagnosis of diseases, composition of drugs tailored towards individual’s genetic structure. In banking organisations, data mining is used to find the loan defaulters, credit scoring and so on. Churn prediction, cross sell and up sell are the benefits witnessed by marketing organisations due to data mining.

Thus, data mining algorithms learns the rules from the large databases which may help the organisation to provide better service, quality at a lower price.

Once the data is released for the purpose of mining, there is every chance of privacy breach. Thus, a technique is required that publishes data while preserving the right balance between individual privacy and data utility thus forming the key area of research. For example, banks might wish to collaborate in order to detect the fraudulent behaviour of customer in its early stages. This requires the banks to share financial records of their customers. In such cases, the banks would like to provide data in a manner that enables the miner to draw inferences without violating the privacy of the individual customer.

In recent years several privacy preserving data mining (PPDM) solutions have been proposed. Generally, used techniques for privacy preserving are distribute subset of data, distribute distorted data and distribute computation. All of these techniques have their own challenges and related issues. There is a clear trade off between accuracy of knowledge gained and the privacy. That is higher the accuracy lower the privacy and lower the accuracy higher the privacy. Hence, PPDM remains as an open research issue.

In this paper, we address issues related to PPDM in particular to privacy preserving data classification. General classification techniques have been extensively studied in the literatures (Mitchell, 1997). The main purpose of classification technique is to build a model (i.e., classifier) to predict the categorical class labels of data tuples based on a training data set where the class label of each record is given. The classifier is usually
represented by classification rules, decision trees, neural networks, or mathematical formulae that can be used for classification. There has been need of designing solutions for privacy preserving data classification models. The objective of privacy preserving data classification is to build accurate classifiers without disclosing private information in the data being mined. The performance of privacy preserving techniques should be analysed and compared in terms of both the privacy protection of individual data and the predictive accuracy of the constructed classifiers (Zhang et al., 2005).

This paper suggests an approach to PPDM, in particular to data classification using the concept of fuzzy sets. It shows that the data privacy can be maintained without compromising the accuracy of the result if features are transformed into fuzzy sets. Reason for the same is that, if data points are transformed into fuzzy sets, the similarity between points is still preserved. The accuracy of the mining results is preserved as the grouping of each attribute is done independently and hence the dissimilarity between points falling in each group is small.

The rest of the paper is organised as follows. Section 2 describes the related work in the area of PPDM. Section 3 gives the outline of our methodology to carry out privacy preserving data classification. Section 4 describes the detailed experimentation and finally we conclude in Section 5.

2 Related works

There has been a surge in recent research activity in the area of PPDM and several papers have been published on various aspects of PPDM. In this section, we discuss some of the previous work closely related to the research presented in this paper and provide references to recent manuscripts that cover the state of the art in the field.

PPDM solution is very much influenced by the research in statistical databases. The main difference in coming up with a solution for privacy preserving between data mining and statistical databases is that, in the case of statistical databases, one knows what data or information is to be preserved (i.e., mean, standard deviations and other statistics). Whereas in the case of data mining, one does not know what will be the information or knowledge extracted from data mining techniques and hence what to be preserved? So in the case of data mining results, one does not know when the privacy has been violated.

Several data masking methods exist based on different assumptions. Data swapping is the commonly used data masking technique. It refers to a method of swapping information from one record to another (Dalenius and Reiss, 1982; Reiss, 1984; Schlorer, 1981). The amount of swapping to be done for the database is dependent on the application and need of organisations. There exist various variants of swapping i.e., records chosen for swapping may be purposely or randomly. Records are chosen purposely for swapping because they are believed to have a greater risk of re-identification. The advantages of swapping are, it is easily implemented and it is one of the best methods of preserving confidentiality. Its main disadvantage is that, even with a very low swapping rate, it can destroy analytic properties, particularly on sub-domains.

Later in 1995, Moore (1995) introduced a controlled way of swapping called rank swapping. In this approach, the values of an individual record are sorted and swapped in a range of k-percent of the total range. A randomisation determines the specific values of record value to be swapped. The procedure is repeated for each variable until all variables have been swapped. The main disadvantage of this approach is the determination of k. If
If \( k \) is relatively small, then analytic distortions on the entire file may be small for simple regression analysis. If \( k \) is large, there is an assumption that the re-identification risk may be reduced (Domingo-Ferrer and Mateo-Sanz, 2001).

Micro-aggregation is another technique for data masking (Defays and Anwar, 1998; Domingo-Ferrer and Mateo-Sanz, 2002). It aggregates the record values of attributes that is intended to reduce re-identification risk. In single ranking micro-aggregation, each attribute is aggregated independently of other attributes. The method is easy to implement. The values of attribute are sorted and divided into groups of size \( k \). In practice, \( k \) is taken to be three or four to reduce analytic distortions. In each group, the values are replaced by an aggregate such as the mean or the median. The micro-aggregation is repeated for each of the attributes that are considered to be usable for re-identification. Domingo-Ferrer and Mateo-Sanz (2001) provided methods for aggregating several attributes simultaneously. The methods can be based on multivariable metrics for clustering variables into the most similar groups. They are not as easily implemented because they can involve sophisticated optimisation algorithms. For computational efficiency, the methods are applied to two to four attributes simultaneously whereas many public use files contain 12 or more attributes. The advantage of the multi-variable aggregation method is that, it provides better protection against re-identification. Its disadvantage is that analytic properties can be severely compromised, particularly if two or three uncorrelated attributes are used in the aggregation. The attributes that are not micro-aggregated may themselves allow re-identification.

Adding noise to the datasets introduced by Kim and Winkler (1995), Kim (1990). Feinberg Fuller (1993) showed that it is theoretically possible to recover means and covariance of a given record on arbitrary sub-domains. Kim (1986, 1990) and Fuller (1993) both showed that the masked dataset by adding noise provides good analytic properties such as regression analysis that closely reproduce regression analysis of the unmasked data. They reasoned that adding noise can yield files with moderate re-identification rates (Fuller 1993; Kim and Winkler, 1995).

In a recent book on PPDM, the editors Aggarwal et al., lists five key directions in the field of PPDM. The five directions mentioned are: privacy preserving data publishing, changing the results of data mining applications to preserve privacy, query auditing, cryptographic methods for distributed privacy and theoretical challenges in high dimensionality (Aggarwal et al., 2008). The work presented in this paper falls into the category of privacy preserving data publishing which is concerned with studying transformation methods associated with privacy such as randomisation and \( k \)-anonymity, application of data perturbation in conjunction with classical data mining methods and studying utility-based methods. Fung et al. (2009) provides a survey on recent developments specifically to the field of privacy preserving data publishing. Atzori et al. (2008) proposes a framework and a formal notion of privacy protection that allows the disclosure of the extracted knowledge while protecting the anonymity of the individuals in the source database. Aggarwal et al. (2008) provides a categorisation of methods such as randomisation and \( k \)-anonymity model that perform some form of transformation on the data for the purpose of privacy preservation. As pointed out by Aggarwal et al., such methods reduce the granularity of representation resulting in a loss of effectiveness of data mining algorithms and the same applies to the fuzzy based transformation technique presented in this paper. A book on PPDM by Vaidya et al. (2006) surveys different classes of PPDM solutions and presents solutions for various data mining tasks like
classification, regression, association rule mining and clustering. Vassilios et al. (2004) propose a classification hierarchy for analysing research work performed in the area of PPDM. Vassilios et al. classify various approaches which have been adopted for PPDM into five broad categories: data distribution, data modification, data mining algorithm, data or rule hiding and privacy preservation. Also, the authors categorise algorithms based on heuristic-based techniques, cryptography-based techniques and reconstruction-based techniques.

In this work, we have presented a privacy preserving technique using the concept of fuzzy set theory, which retains both the privacy as well as the accuracy of result. The current approach is similar to micro-aggregation technique. In the current approach, the individual privacy is retained as reconstruction of original data from transformed data set is not possible.

3 Fuzzy clustering algorithm for PPDM

Fuzzy set has dominated the field of knowledge representation and interpretation for a long time. Fields like machine intelligence, artificial intelligence, data mining and pattern recognition has widely accepted the theory of fuzzy sets in their application areas.

Fuzzy set theory is an aid to represent uncertainty, possibility and approximation. If something is fuzzy, it means that it is difficult to define precisely its boundaries. Human reasoning is not only able to make decisions and classify situations based on partial or ambiguous information; it is actually inclined to it. Fuzzy sets try to mimic what is natural to human reasoning, allowing the imprecise definition of concepts.

Transforming the given set of data points from the crisp set to fuzzy set is known as fuzzification. By the process of fuzzification, each data point is associated with certain degree of membership to every fuzzy set. Every data point is associated with a linguistic variable. A linguistic variable represents a concept that is measurable in some way, either objectively or subjectively, like temperature or age. For each linguistic variable, it should be assigned a set of linguistic terms (values) that subjectively describe the variable. Each linguistic term is a fuzzy set and has its own membership function. It is expected that for a linguistic variable to be useful, the union of the support of the linguistic terms cover its entire domain.

Categorical variables are like facial features. There may be many individuals who share the same features. Numerical variables are like finger prints. When there are a number of numerical values, it is extremely unlikely that two individuals have exactly the same set of values for all variables. Hence, in the current work we have used fuzzy c-means algorithm to fuzzify the numerical attributes of the data set. With the help of fuzzy c-means algorithm each attribute is divided into $c$ fuzzy sets where each data point has some degree of membership with every fuzzy set.

J. C. Bezdek (1981) introduced the fuzzy c-means algorithm for clustering the set of input data points into fuzzy clusters. The fuzzy c-means algorithm allows each input data point to belong to $c$ clusters with a membership value lying between zero and one. Fuzzy c-means work on the principle of minimising the weight to the total weighted mean square error. Details of the fuzzy c-means algorithm can be found in the literatures Bezdek (1981).

Once the dataset is fuzzified, classification of the dataset is done. For classification, only the independent variables are fuzzified and dependent value is left as it is. The
records in fuzzified database are stored with the linguistic term for each attribute value with its corresponding class label. The data miner can generate the decision rules using the linguistic records. Since the exact value of individual records is not given to the data miner, there is no threat to the privacy. However, the discretisation has been done using fuzzy c-means algorithm, which discretises the database using the distance function, hence, the similarity across the records are preserved leading to accurate mining results.

3.1 An illustrative example

Let $D$ be a dataset consisting of $N$ tuples (instances or rows). Each tuple is denoted by $t_i$ where $i$ (tuple index) varies from 0 to N-1. Let $M$ be the total number of attributes or features (column) in the dataset. Each attribute or feature is denoted as $f_j$ where $j$ (feature index) varies from 0 to M-1. Let $V_{ij}$ denote the value present at row $i$ and column $j$. The total number of values in the dataset will be $N*M$. The proposed algorithm is a two-step process that consists of the following sequential operations:

Step 1 Independent fuzzification of numeric features

Let $f^c$ and $f^n$ denote the set of features having categorical and numeric values. Let FCM() denote a fuzzy c-means function that takes a numeric feature as input and returns a transformed or fuzzified feature. The first step in the solution approach consists of applying fuzzy c-means function to each of the numeric features. The fuzzy c-means procedure is applied to each numeric attribute independently as a result of which a transformed feature set $f^{\text{nc}}_n$ is obtained. The presence of $t$ in the set $f^{\text{nc}}_n$ indicates a transformed feature set. The categorical feature set remains as it is as the fuzzy c-means procedure is applied only to the numeric features. The operation performed by the first step can be denoted mathematically as:

$$f^{\text{nc}}_n = \text{FCM}(f^n)$$

As a result of this step, the original dataset $D$ is transformed to a new dataset denotes by $D^t$. The feature set in dataset $D$ consists of sets $f^c$ and $f^n$ whereas the feature set in dataset $D^t$ consists of sets $f^c$ and $f^{\text{nc}}$. The number of rows and columns in the transformed dataset remains the same as in original dataset $D$, i.e., the number of rows is equal to $N$ and the number of columns is equal to $M$. Each value in the numeric feature of the transformed dataset is a fuzzy value having a degree of membership to each of the ‘c’ fuzzy-sets in which the numeric feature is partitioned. The degree of membership of each value lies in the continuous interval (0,1) such that the sum of the degree of membership is equal to 1.

Step 2 Assigning a single class or category to a fuzzy-value (denoted by a vector having degree of membership across various fuzzy-sets)

As a result of Step 1, each value in the numeric feature belongs to different fuzzy-sets having a degree of membership to each fuzzy-set in the range of zero and one. Thus each value can be viewed as a vector rather than a single scalar or categorical value. Classification algorithms cannot be applied directly on such dataset. Hence, the next step consists of converting or replacing the vector (degree of membership to each fuzzy set) to a single value. This is done by taking the fuzzy-set for which the value has the highest
degree of membership. Subsequent to the above operations, classification algorithms can
be applied to perform the data mining tasks while keeping the individual privacy
preserved.

Table 1 Sample dataset

<table>
<thead>
<tr>
<th>SSN no</th>
<th>Name</th>
<th>Age</th>
<th>Annual income (US$)</th>
<th>Car owner</th>
<th>House owner</th>
<th>Customer category</th>
</tr>
</thead>
</table>
| 52     |      | 83K | Yes                 | Yes       | Platinum
| 34     |      | 91K | Yes                 | Yes       | Platinum
| 61     |      | 47K | Yes                 | Yes       | Platinum
| 73     |      | 85K | Yes                 | No        | Platinum
| 48     |      | 52K | Yes                 | Yes       | Gold         |
| 44     |      | 86K | Yes                 | No        | Gold         |
| 24     |      | 55K | No                  | No        | Gold         |
| 51     |      | 51K | Yes                 | No        | Gold         |
| 29     |      | 24K | No                  | No        | Silver       |
| 36     |      | 39K | No                  | No        | Silver       |
| 58     |      | 47K | Yes                 | No        | Silver       |
| 40     |      | 26K | Yes                 | No        | Silver       |

Table 2 Fuzzified dataset

<table>
<thead>
<tr>
<th>SSN no</th>
<th>Name</th>
<th>Age</th>
<th>Annual income (US$)</th>
<th>Car owner</th>
<th>House owner</th>
<th>Customer category</th>
</tr>
</thead>
</table>
| Old    | High  | Yes  | Yes                 | Yes       | Platinum
| Young  | High  | Yes  | Yes                 | Yes       | Platinum
| Old    | Medium| Yes  | Yes                 | No        | Platinum
| Old    | High  | Yes  | Yes                 | No        | Gold         |
| Old    | Medium| Yes  | Yes                 | No        | Gold         |
| Young  | Medium| No   | No                  | No        | Silver       |
| Old    | Medium| Yes  | No                  | No        | Silver       |
| Young  | Low   | No   | No                  | No        | Silver       |
| Old    | Medium| Yes  | No                  | No        | Silver       |
| Young  | Low   | Yes  | No                  | No        | Silver       |

Notes: If ((Income = High) and (Age = Young) and (Home = Yes)) $\rightarrow$ Platinum
If ((Income = High) and (Age = Young) and (Home = No)) $\rightarrow$ Gold
If ((Income = High) and (Age = old)) $\rightarrow$ Platinum
If ((Income = Medium) and (Home = Yes)) $\rightarrow$ Platinum
If ((Income = Medium) and (Home = No)) $\rightarrow$ Gold,
If ((Income = Low)) $\rightarrow$ Silver
Table 1 shows a sample dataset of 12 customers belonging to different types of credit card. The dataset consists of 12 records and seven columns. The two columns SSN no and name have been masked as they uniquely identify individuals. Age and annual income are the attributes which take numerical value as input whereas attributes car owner and house owner takes categorical values as input. The last attribute i.e., customer category is the class variable and takes categorical values.

As the starting step, all the numerical variables are fuzzified independently using fuzzy c-means algorithm. We clustered the attribute age into two fuzzy clusters whereas the attribute annual income was clustered into three fuzzy clusters. The value of age attribute was replaced by its cluster name such as young and old. Similarly, the value of annual income attribute was replaced by its cluster names, high, medium and low. Thus the fuzzified database was built. Table 2 shows the final fuzzified dataset. As can be observed from the Table 2, the individual privacy is preserved.

Now we build a tree on the fuzzified dataset. The tree generated from fuzzified dataset is shown in Figure 1. Rules formed due to the tree are as follows:

![Figure 1](image)

The confusion matrix for the fuzzified sample dataset is shown in Table 3. As can be seen from the table the entire platinum card holders are classified accurately whereas there is a misclassification in the case of gold and silver card holders. This can be due to the reason that the gold and silver card holders are not well linearly separable and hence a large value for number of fuzzy cluster may be able to classify these classes more accurately.

<table>
<thead>
<tr>
<th></th>
<th>Gold</th>
<th>Silver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platinum</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Gold</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Silver</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
4 Experimental results and discussions

This section briefs about the experiments performed for PPDM. Fuzzy c-means algorithm is implemented in Java 1.4 whereas Naive Bayes and J48 classification algorithms were taken from open source WEKA software available at Weka. Experiments were conducted over three datasets namely, Iris, Image Segmentation and Australian Credit datasets obtained from WEKA.

Iris dataset consists of 150 instances with four numerical attributes. These 150 instances belong to three classes namely, Iris setosa, Iris versicolor and Iris virginica. Segment Challenge dataset is a subset of the segmentation data (i.e., a subset of the combined original training and test datasets). The selected dataset has 1,500 instances. The distribution of classes is not stratified in this selected subset. The dataset has 19 continuous attributes. Each instance may belong to any one from the seven classes namely, brickface, sky, foliage, cement, window, path and grass. Australian Credit dataset consists of 690 instances with 14 attributes. Out of these 14 attributes, six are numerical and eight are categorical. All these instances belong to two classes namely, good and bad. For all the datasets, in our experiments, we used only numeric attributes. In the current work, we concentrated on the effect of the mining results due to fuzzification.

In our experiments, we have used ID3 and Naive Bayes classifiers to study the effect of fuzzification on the classifier models. ID3 builds a decision tree from the training dataset. The resulting tree is used to classify the samples of test dataset. The training dataset has several attributes and belongs to one class. The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. ID3 uses information gain to help it decide which attribute goes into a decision node. The advantage of learning a decision tree is that a program, rather than knowledge engineer, elicits knowledge from an expert (Mitchell, 1997; Quinlan, 1993). Naive Bayes classifier is a simple but efficient classifier. Naive Bayes is based on a Bayesian formulation of the classification problem which uses the simplifying assumption of attribute independence. It is simple to implement and gives accurate results (Mitchell, 1997).

The classifier accuracy was evaluated using cross validation technique. Cross validation is one of the approaches to see how well the model performs on test data. In our experiments, we reported the accuracies of the classifier using $k$-fold cross validation technique (Mitchell, 1997). In $k$-fold cross validation approach, the dataset is divided into $k$ subsets. One from the $k$ subsets is used as the test set rest $k-1$ subsets are used to train the model combined. The process of selecting the test set and training set is repeated for $k$ times thus ensuring that the model is trained as well as tested over all the records at least once. The average accuracy of the $k$ runs is taken as the classifier accuracy. For our experiments $k$ was chosen to be ten.

Figure 2 and 3 shows the graph between number of clusters and the accuracy of the classifier. From the figures, it can be observed that as the number of clusters increases the accuracy of the classifier also increases.

We also made a study over the misclassification rate and correct classification rate for both the classifiers. Correct classification rate is the ratio of the number of test records correctly assigned to the total number of records in test set. The misclassification rate is
defined as the ratio of number of test records misclassified to the total number of records in training dataset. We report in Table 4 and 5 the experimental results for three datasets. In our experiments we recorded the false positive rate and true positive rate for different number of clusters for the two classifiers Naive Bayes and ID3. As can be observed from the tables that, with the increasing number of cluster the misclassification rate is decreasing whereas the correctly classification rate is increasing. The same results are observed for all the three datasets for both the classifiers. In the case of Australian Credit data set and Image Segment dataset with Naive Bayes classifier, we were able to classify all the records correctly with small number of clusters.

Figure 2  Graph between number of clusters and ID3 classifier accuracy

Figure 3  Graph between number of clusters and Naive Bayes classifier accuracy
Fuzzy based clustering algorithm for privacy preserving data mining

Table 4  Table showing true positives and false positives for Naive Bayes classifier

<table>
<thead>
<tr>
<th>No of clusters</th>
<th>Iris</th>
<th>Image Segmentation</th>
<th>Australian Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCR</td>
<td>MCR</td>
<td>CCR</td>
</tr>
<tr>
<td>2</td>
<td>0.61</td>
<td>0.38</td>
<td>0.47</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.22</td>
<td>0.62</td>
</tr>
<tr>
<td>4</td>
<td>0.80</td>
<td>0.20</td>
<td>0.65</td>
</tr>
<tr>
<td>5</td>
<td>0.81</td>
<td>0.18</td>
<td>0.68</td>
</tr>
<tr>
<td>6</td>
<td>0.80</td>
<td>0.20</td>
<td>0.71</td>
</tr>
<tr>
<td>7</td>
<td>0.83</td>
<td>0.16</td>
<td>0.76</td>
</tr>
<tr>
<td>8</td>
<td>0.84</td>
<td>0.16</td>
<td>0.75</td>
</tr>
<tr>
<td>9</td>
<td>0.80</td>
<td>0.20</td>
<td>0.78</td>
</tr>
<tr>
<td>10</td>
<td>0.87</td>
<td>0.12</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Notes: CCR = correctly classification rate, MCR = misclassification rate.

Table 5  Table showing true positives and false positives for ID3 Classifier

<table>
<thead>
<tr>
<th>No of clusters</th>
<th>Iris</th>
<th>Image Segmentation</th>
<th>Australian Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCR</td>
<td>MCR</td>
<td>CCR</td>
</tr>
<tr>
<td>2</td>
<td>0.62</td>
<td>0.38</td>
<td>0.81</td>
</tr>
<tr>
<td>3</td>
<td>0.78</td>
<td>0.21</td>
<td>0.96</td>
</tr>
<tr>
<td>4</td>
<td>0.82</td>
<td>0.18</td>
<td>0.96</td>
</tr>
<tr>
<td>5</td>
<td>0.86</td>
<td>0.14</td>
<td>0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.87</td>
<td>0.12</td>
<td>0.99</td>
</tr>
<tr>
<td>7</td>
<td>0.89</td>
<td>0.10</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.91</td>
<td>0.08</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0.91</td>
<td>0.08</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.89</td>
<td>0.10</td>
<td>1</td>
</tr>
</tbody>
</table>

Notes: CCR = correctly classification rate, MCR = misclassification rate.

We feel that the combination of different number of clusters for different attributes may result in better classifier accuracy. The classifier accuracy dropped significantly even with small number of outliers. This is due to the choice of number of clusters in the fuzzy c-means algorithm. Hence a technique should be devised for clustering all outliers or noise to one cluster using kernel functions or any technique which can handle such situation. In our ongoing research this forms the main concern.

With the ID3 classifier both Australian Credit and Image Segmentation datasets, we were able to achieve the misclassification rate of zero. However, we were not able to achieve the misclassification rate of zero for Iris dataset. The possible reason for this is in case of Iris dataset, class 2 and class 3 are not linearly separable. In the case of Naive Bayes classifier, the correctly classification rate recorded was more than 0.85% for all the three datasets with more than ten clusters. This suggests that as the number of clusters increases, the better classifier accuracy can be achieved.

The transformation of the attributes in the database to fuzzy dimension is a promising solution for preserving privacy since exact values of the attributes cannot be revealed.
from the linguistic terms. The transformation of the numerical attribute to the linguistic attribute is irreversible process, privacy of individual is totally guaranteed. For example, when value of a record for attribute income is changed from 35K US dollars to low, the miner comes to know of the meaning but the exact value 35K US dollars can not be predicted exactly. The miner may make an arbitrarily guess that might be correct, which happens very rarely and it can be just considered as a chance of guessing correctly. The transformation not only preserves the similarities between objects but also the data transformed retains its meaning.

With the fuzzy clustering approach, each object is allowed to belong to more than one cluster, thus relaxing the constraint of having at least $k$ elements is relaxed. This results into more compact clusters. These records may contribute to the aggregates of several clusters. Also, several clusters can be used for masking a particular object decreasing the disclosure risk of the object Domingo-Ferrer and Vicen (2003).

Above results and discussions shows that the fuzzy based data transformation provides accurate results in terms of classification problem. The database owner has full privacy with its database and hence can share the fuzzified database for mining tasks. The above results suggests that the fuzzy based data transformation can be adopted for PPDM in particular to classification tasks

5 Conclusions

Data mining is today’s business need for providing better service, quality at low price. Data mining makes the sensitive data exposed to the unauthorised person. In this paper, we have addressed the issue of PPDM using fuzzy based technique where the data is discretised into different linguistic variables. The current approach has been adopted for classification task. In our experiments, we adopted two well-known classification algorithms namely, Naive Bayes and ID3. A study was made over the classifier accuracy. Experiments were performed over three datasets. These experiments suggest that the fuzzy based technique can be adapted for PPDM tasks in particular to data classification.

In our ongoing work, we are exploring the relationship between $m$ (fuzzifying parameter in Fuzzy c-means clustering algorithm) and mining results. We are also looking to extend the current approach on various other classifiers. The framework can be extended and tested on various other data mining tasks such as association rule mining, clustering and sequential pattern mining. The work presented in this paper can also be extended to categorical attributes with modification over fuzzy c-means clustering algorithm.

Acknowledgements

The part of the work was carried out, when the first and third author were associated with SETLabs, Infosys Technologies Limited, Bangalore, India.
References


