This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30044586

Reproduced with the kind permissions of the copyright owner.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright : 2008, IEEE
Learning in Imbalanced Relational Data

Amal S. Ghanem, Svetha Venkatesh, Geoff West
Department of Computing, Curtin University of Technology
GPO Box U1987, Perth, 6845, Western Australia
amalsaleh.ghanim@postgrad.curtin.edu.au, (s.venkatesh, g.west)@curtin.edu.au

Abstract

Traditional learning techniques learn from flat data files with the assumption that each class has a similar number of examples. However, the majority of real-world data are stored as relational systems with imbalanced data distribution, where one class of data is over-represented as compared with other classes. We propose to extend a relational learning technique called Probabilistic Relational Models (PRMs) to deal with the imbalanced class problem. We address learning from imbalanced relational data using an ensemble of PRMs and propose a new model: the PRMs-IM. We show the performance of PRMs-IM on a real university relational database to identify students at risk.

1 Introduction

In many universities, the retention rates at the end of the first-year program are regarded as a key performance indicator, and an area in which considerable resources are invested to improve learning outcomes. One of the methods used to improve the retention rates is modelling the students' information stored in the university relational database to investigate the relationships between the students' performances in their units.

Learning a student model from a university relational database includes: investigating a large number of records and relationships representing the students' information and the performances in their units, and at the same time considering the imbalanced class problem that exists in the data distribution of students' performances, where the number of students at risk ('Fail' students) is small compared to the other classes.

Although different learning techniques, such as regression analysis and Bayesian Networks (BNs), have been successfully applied in a number of student modelling applications [4, 14]; these traditional techniques are still inadequate in handling complex relational domains. These techniques require all the domain data to be presented in a single flat file of fixed variables and their values, whereas most real-world datasets are stored as relational databases that consist of a collection of tables and relationships.

The imbalanced class problem is common in real-world applications, and there is a rich literature of imbalanced learning techniques applied to single flat files [2, 3, 11, 12]. In contrast, only few attempts are proposed to handle the imbalanced problem in relational domains [7, 9, 13]. However, none of these methods have used a specifically designed relational learning algorithm, except the work proposed by Sen and Getoor [13] that uses cost-sensitive learning, but in turn this method includes the challenge of assigning the proper misclassification cost.

In this paper, we propose an approach to deal with the challenges of student modelling from imbalanced relational data. In terms of learning in relational data, we adapt a probabilistic technique designed specifically for relational data called Probabilistic Relational Model (PRM) [5, 8]. PRM can be learned directly from the university dataset and can be used with great flexibility to model individual students and answer queries related to a student's performance. In terms of the imbalanced issue, we build an ensemble of PRM models called PRMs-IM for the two-class case, where each PRM model is trained on a balanced subset of the data. Each balanced subset incorporates all the samples from the minority class and a random selection of equal samples from the majority class. Then, these models will be combined using a weighting voting strategy. The novelty is in learning PRM from real university relational datasets and handling the imbalanced class problem.

2 Related work

PRMs. Probabilistic Relational Models (PRMs) [5, 8] have been designed specifically for relational learning and inference. PRMs can be viewed as an exten-
cision of Bayesian Networks (BNs) to incorporate the relational structure by specifying a template for a probability distribution over a relational database. The template consists of the relational database of the domain, and the probabilistic schema that describes the dependencies between the attributes of the domain. 

The input of the PRM consists of the domain relational database consisting of a set of tables and relationships. Each table includes a set of attributes, and each attribute takes on values from a fixed domain. The PRM probabilistic schema consists of: the dependency structure, which is a directed acyclic graph defining the interactions between the domain variables, and the structure parameters that consists of the conditional probability distributions of the variables. Given the learned dependency model and parameters, the PRM model can be used to answer different queries about a new instance by deriving a customized network [5]. PRMs have been successfully applied to real-world applications: databases selectivity estimation [6] and student modelling in virtual laboratories [10]. In our experiment we use the PRM learning and inference techniques described in [5].

**Imbalanced class problem.** Traditional learning algorithms such as Bayesian networks and decision trees perform poorly on imbalanced class distribution [9], as the algorithms get biased towards the majority class resulting in poor prediction for the minority. There is a rich literature of imbalanced class learning techniques applied to single flat files, including: biasing the learning algorithm towards the minority class [3, 11], re-sampling the data distribution by over-sampling or under-sampling [2], and trying to minimize the cost errors by assigning high misclassification costs to the minority class (cost-sensitive learning) [12].

However, only few attempts have been proposed to handle the imbalanced relational problem, examples include: g-mean decision trees [9], cost-sensitive learning for structured data [13], and the MVC-IM algorithm [7] that combines the learning from multiple flat views of the dataset.

### 3 Methodology

In contrast to the existing relational imbalanced methods, in our proposed approach (PRMs-IM), we employ PRMs that are designed specifically to learn from relational databases without the need to use decision trees as in [9], or converting into flat views as in the MVC-IM algorithm [7]. Furthermore, we aim to take advantage of sampling and insensitive learning techniques for relational data and thus avoid the challenge of setting the proper misclassification costs such as proposed by Sen and Getoor [13].

In the imbalanced situation of a majority class $C_{mj}$ with samples $n_{mj}$ and a minority class $C_{mr}$ with samples $n_{mr}$, we extend PRMs to deal with this situation by partitioning the $n_{mj}$ based on the statistical distribution of the dataset, in a similar approach to that discussed in [1, 11] on flat files. The idea is to use an ensemble of a set of PRM components, in which each component of the ensemble represents a PRM model over an individual balanced relational subset. Each subset would include all the $n_{mr}$ and an equal number of the $n_{mj}$. The selection of the $n_{mj}$ can be performed randomly with or without replacement. The number of subsets is determined as the difference between the number of the $n_{mj}$ and that of the $n_{mr}$. For example, for a $C_{mj}$ that is four times the $C_{mr}$, then four subsets would be created.

The components are then combined using the weighted voting strategy [1]. In this strategy, each component has a different weight affecting the final classification result. For example, for a component $Ci$ with data subset $Si$, the corresponding weight $Ci_{w}$ is calculated as the average of the performance accuracy resulting from testing $Ci$ on the other components’ datasets.

For a testing sample $x$, each $Ci$ outputs the probability scores $(Ci(x)_{mj}, Ci(x)_{mr})$ for assigning $x$ to the $C_{mj}$ and $C_{mr}$, respectively. Then, the score of each class is calculated as the summation of the weighted scores of the components. Therefore, $x$ is classified to the class of the largest weighted probability score, i.e. $F(x) = \arg \max_{m \in \{mr,mj\}} (\sum_{ci} Ci(x)_{m} * Ci_{w})$.

However, for relational databases, each data subset needs to be a balanced representative of the relational data. Therefore, PRMs-IM applies a customized approach of the ensemble method. For $n$ PRMs components, we first create $n$ new empty databases. Each of the datasets will be allocated first with all the $n_{mr}$. We then randomly allocate the $n_{mj}$ to the $n$ datasets, keeping in mind that the selected number of $n_{mj}$ should not exceed the $n_{mr}$. Finally, we add all the related records from the other tables that are linked to those records in the target table. This procedure creates $n$ roughly balanced sub-databases, which include all the $n_{mr}$ and a random selection of equal samples from the $C_{mj}$.

### 4 Experiments

**Databases:** We use the Curtin University student database for students enrolled in the Bachelor of Computer Science (BCS) and Bachelor of Commerce (BCom). The data consist of the following tables: PersonInfo, AcademicInfo, and a number of tables representing the first year units taken in semester I and II.
The Personal_Info table includes the attributes: age, gender, is_international, and is_English_home_language, which take on values {16-19, 20-29, 30-40}, {Male, Female}, {Yes, No}, {Yes, No}, respectively. The Academic_Info table includes: Preference_no representing the student’s preference of study, which takes on values {1, 2, 3, 4}. The tables of semester I units include: grade that takes values of {F, 5, 6, 7, 8, 9} representing the grades categories of {0-49, 50-59, 60-69, 70-79, 80-89, 90-100}. The tables of semester II units include: status attribute with values of {Pass, Fail}. For the BCS dataset, semester I units are: ST151/prerequisite for ST152, Maths101, FCS151/prerequisite for FCS152, and English101. Semester II units are: ST152, FCS152 and IPE151. As for the BCom, the units of semester I are: BIS100, ACCT100, LFW100 and ECON100. Semester II units are: MGT100 and MKT100.

Our model uses a two-class problem to predict the performances (Pass or Fail) of students in semester II units given their information and grades in semester I. In particular, we are more interested in correctly detecting the students at risk. As each of the semester II units has a different number of minority and majority samples, the number of PRM components and the data partitions differ between the units even within the same degree. Table 1 shows a summary of the training datasets for students enrolled in the period 1999-2005. Each of these datasets represents a separate experiment to predict the performance of the target unit. For each dataset, 5-cross validation was employed. Furthermore, the data of 2006 students was kept separately as an independent testing set. The BCS dataset is restricted by the prerequisite units, thus the training set includes only the set of students who passed the prerequisites and enrolled in semester II units, which resulted in a low number of students in the BCS dataset.

Experimental results: The results obtained from PRMs-IM on the databases are shown in Table 1. The results are presented in comparison to three classification algorithms: (1) the traditional classification achieved by applying BN and PRMs on the imbalanced dataset, (2) MVC-IM [7] (using Naive Bayes), the multi-relational algorithm for imbalanced datasets, as this algorithm has a similar approach of combining classifiers but differs in that each classifier focuses on a subset of features, and (3) APRI [3] that biases the BN classifier in favour of the target attribute, as it extends the BN in terms of handling imbalanced data but only on flat files.

PRMs-IM is evaluated using the measurement metrics usually used for imbalanced classification algorithms: the Receiver Operating Characteristics (ROC) curves and the Area under ROC (AUC), as used by [3, 7]. The ROC curve visualizes the trade off between the false positive rate and the true positive rate. To compare several models using the ROC curves, the AUC is used to get a single value of the classifier performance. The higher the AUC value, the better the classifier. Tables 2 and 3 present the AUC results obtained from each of the experiments using 5-fold cross validation and testing on 2006-students, respectively. The best results for each dataset are shown in bold. A sample of corresponding ROC curves is shown in Figures 1 and 2.

The results from the ROC curves and AUC show the poor performance of applying the traditional techniques: BN and PRMs, directly to the imbalanced dataset. The results also show that PRMs-IM was able generally to improve over all the other techniques. Exceptions were in the cases of the ST152 and MKT100 datasets, in which MVC-IM scored slightly better in the first and equalled the score in the latter dataset. This is the subject of further investigation.

### Table 1. Summary of the dataset used.

<table>
<thead>
<tr>
<th>Degree</th>
<th>Dataset</th>
<th>No. Samples</th>
<th>No. PRM Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Fail</td>
<td>Pass</td>
</tr>
<tr>
<td>BCom</td>
<td>MGT100</td>
<td>159</td>
<td>1556</td>
</tr>
<tr>
<td></td>
<td>MKT100</td>
<td>88</td>
<td>1627</td>
</tr>
<tr>
<td>BCS</td>
<td>ST152</td>
<td>12</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>FCS152</td>
<td>11</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>IPE151</td>
<td>7</td>
<td>63</td>
</tr>
</tbody>
</table>

### Table 2. AUC results for 5-fold cross validation.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Dataset</th>
<th>BN</th>
<th>PRMs</th>
<th>APRI</th>
<th>MVC-IM</th>
<th>PRMs-IM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MGT100</td>
<td>0.488</td>
<td>0.558</td>
<td>0.737</td>
<td>0.871</td>
<td>0.914</td>
</tr>
<tr>
<td></td>
<td>MKT100</td>
<td>0.440</td>
<td>0.508</td>
<td>0.735</td>
<td>0.846</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>ST152</td>
<td>0.365</td>
<td>0.300</td>
<td>0.758</td>
<td>0.849</td>
<td>0.849</td>
</tr>
<tr>
<td></td>
<td>FCS152</td>
<td>0.365</td>
<td>0.380</td>
<td>0.846</td>
<td>0.711</td>
<td>0.901</td>
</tr>
<tr>
<td></td>
<td>IPE151</td>
<td>0.704</td>
<td>0.771</td>
<td>0.889</td>
<td>0.863</td>
<td>0.913</td>
</tr>
</tbody>
</table>

### Table 3. AUC results for the 2006-testing set.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Dataset</th>
<th>BN</th>
<th>PRMs</th>
<th>APRI</th>
<th>MVC-IM</th>
<th>PRMs-IM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MGT100</td>
<td>0.413</td>
<td>0.408</td>
<td>0.645</td>
<td>0.845</td>
<td>0.921</td>
</tr>
<tr>
<td></td>
<td>MKT100</td>
<td>0.603</td>
<td>0.572</td>
<td>0.714</td>
<td>0.704</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>ST152</td>
<td>0.187</td>
<td>0.125</td>
<td>0.913</td>
<td>0.937</td>
<td>0.875</td>
</tr>
<tr>
<td></td>
<td>FCS152</td>
<td>0.400</td>
<td>0.380</td>
<td>0.840</td>
<td>0.730</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td>IPE151</td>
<td>0.571</td>
<td>0.59</td>
<td>0.881</td>
<td>0.863</td>
<td>0.954</td>
</tr>
</tbody>
</table>
5 Conclusion

We outline a framework (PRMs-IM) for extending PRMs to handle the relational imbalanced class problem. PRMs-IM was applied to an undergraduate imbalanced relational dataset as an ensemble of PRMs, and achieved promising results in identifying students at risk in computing and commerce degrees. The results of PRMs-IM improved dramatically over those of standard PRMs and BNs algorithms and compared favorably with special algorithms dealing with imbalanced relational data.

References