USING CLASSIFICATION ON UPISI 2011 DATABASE

Kiril Kiroski
Institute of Informatics, Faculty of Natural Sciences and Mathematics

Magdalena Kostoska
Institute of Informatics, Faculty of Natural Sciences and Mathematics

Ana Madevska Bogdanova
Institute of Informatics, Faculty of Natural Sciences and Mathematics

ABSTRACT

Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. [1] Classification is one of the most common techniques of data mining, which occurs very frequently in everyday life. Classification is the central data mining technique that we use in this research. Since classification involves diving up objects so that each of these objects will fall into one of mutually exhaustive and exclusive categories we call classes, we use this technique to classify the numbers of enrolment of a student needed to complete a certain course. In this paper, we will be using some of the most frequently used classification methods. These methods will be tested on a chosen dataset to serve as an example of which of these methods is most suitable for such dataset form. The dataset is extracted from Application "Upisi", and treated with three classification approaches: Naive Bayes, Nearest Neighbour and Decision Trees. Our main goal is to compare the results gain from the database of the application Upisi in 2010 [2] and the results from 2011. Through this comparison we establish how to improve our classification technique.

I. INTRODUCTION

- "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand et al. [3]).
- "Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" (Evangelos Simoudis in Cabena et al. [4])

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining today is not only used in science circles, but everywhere we have big collections of data we can use to get some insight into their meaning. For example, almost every bigger sport collective uses data mining to discover where their strengths and weaknesses are, they use it to discover more information about their opponents, and find a way to play according to this newly acquired knowledge. It has been used for many years by businesses, scientists and governments to sift through volumes of data such as airline passenger trip records, census data and supermarket scanner data to produce market research reports.[5]

In this paper we will present results from our classification on the database used for the application Upisi, which is used by Institute of Informatics as primary student information keeping application. [2] The goal of this research is to predict the number of enrolments a student needs to complete a given course, given the numbers of enrolment to previous dependable courses. In the year 2010 we already conducted this kind of classification, and the main point of this research is to confirm or to improve the model, since new amount of data is now available in the system, according to the new course enrolments for the summer semester in 2011. We will also compare results using newly produced data versus data used during last year's experiments.

As a software tool for this research we used Weka. [9] All results provided in this paper were product of experiments conducted using this tool on the newest data in the database. There are more experiments conducted in the last year beyond the scope of this paper, and we will give brief overview of them, as directions for future research.

II. DATA PREPARATION

When beginning work on a data mining problem, it is first necessary to bring all the data together into one, unified set of instances. This set is also referred to as dataset, and it represents form of data suitable for use with various data mining techniques. There are more models for data mining, which can be used as blueprints organizing the process of gathering and analyzing data, disseminating and implementing results and monitoring improvements. [6]

First we introduce you to our dataset, which is described in table 1.

Table 1: Description of the dataset

Attribute	**Type**	**Values**
Code	Nominal	YYYY-{3,4}G-XX
Part time	Nominal	{t,f}
Prog. Basics - Grade	Numeric	5 – 10
Prog. Basics – Num. enrl.	Numeric	1 – 4
Obj.&Vis. Prog. – Grade	Numeric	5 – 10
Obj.&Vis. Prog– Num. enrl.	Numeric	1 – 4
Comp. Arch. - Grade	Numeric	5 – 10
Comp. Arch. – Num. enrl.	Numeric	1 – 4
Data Struc. - Grade	Numeric	5 – 10
Data Struc. – Num. enrl.	Numeric	1 – 4
Databases- Grade	Numeric	5 – 10
Databases – Num. enrl.	Numeric	1 – 4

In the dataset we use, we first pre-process our data to remove incomplete entries, obsolete or redundant fields and outliers, if there are any. Since the data we are using are intended to be used by a specific application, we first have to transform data into more suitable form for our analysis, consistent with our policy. We use different approach in data cleaning from the year 2010.

III. CLASSIFICATION

Classification is a task that occurs very frequently in everyday life. Essentially it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as classes. The term 'mutually exhaustive and exclusive' simply means that each object must be assigned to precisely one class, i.e. never to more than one and never to no class at all. [7]

Classification is the process of learning a function that maps (classifies) a data item into one of several predefined classes. This data item will be the target for classification process, and it will contain data interesting for the researchers and produce knowledge from this data. The data mining model examines a large set of records, each record containing information on the target variable as well as a set of input or predictor variables. Researcher should be able to classify data records which are still not in the dataset, on base of the characteristics associated with the predictor variables.

Classification methods require thorough data preparation, so as to be able to give more precise prediction. Then, the data set containing predictor variables and target variable is examined, which is the way the algorithm (software) "learns" the ground rules how predictor variables are associated with the target classes. The data set on which learning is performed, is called *training set.* When these rules are set, algorithm is ready to proceed with assessing the accuracy of the classification procedure. This assessment is performed through using established classification on the *test data*. Accuracy of the classification is given with the following equation:

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}} \quad (1)$$

For the purposes of this analysis, we will use following three classification methods:

Naive Bayes
k-nearest neighbor
Decision Tree

In the research conducted in 2010 as a class we used the attribute *Number of enrolment of course Databases* with the following values: [2]

One – which means that the student will have only one enrolment of the course

Two – which means that the student will have two enrolments of the course

Three – which means that the student will have three enrolments of the course

Four – which means that the student will have four enrolments of the course

Since then, a new amount of data has been made available in the system, due to course enrolments for the summer semester in 2011. To be able to address the new state of the system, our class model had to change. Also a new situation occurred – we found big amount of data for students that haven't passed certain course yet. According to the size of the available data we decided not to leave this data unused. So we created the following values for the class attribute:

One enrollment
Two enrolments
Three enrolments
} Passed = Y

At least three enrolments
At least two enrolments
At least three enrolments
} Passed = N

Out dataset for experiments with these new classes contains 530 instances, compared to 412 instances used in experiments conducted last year.

A. Naïve Bayes

Naive Bayes is a classification method which uses no rules, decision tree or some other explicit classifier representation. Instead, Naive Bayes uses *probability theory* to find the classification with greatest likelihood.

The probability of this "class" can be a value between 0 (impossible) and 1 (certain), and it is improved with a longer series of trials. Since we are not interested in only one class, but in a set of alternative classes, they must be *mutually exclusive* and *exhaustive*, so one and only one will always occur. The probability model for the classifier is the following conditional model:

$$p(C|F1,\dots,Fn) \quad (2)$$

for the dependent class variable C with small number of outcomes (classes), conditional on variables F1 through Fn. Using Bayes' theorem, it can be written as:

$$p(C|F1,\dots,Fn) = \frac{p(C)p(F1,\dots,Fn|C)}{p(F1,\dots,Fn)} \quad (3)$$

This equation can be written also as:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \quad (4)$$

Using Naive Bayes Algorithms in Weka is possible only with categorized values. During experiments we worked with several Naive Bayes methods, including AODE, AODEsr, HNB, NaiveBayes, BayesNet, etc. We used 8 predictor fields (number of times course is attended and grade achieved for every course preceding BP in the chain - OP, OVP, AK and SPA). Results (the confusion matrix and summary) are shown on Figure 1 and Figure 2.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 318   0   0   0  20   0 |   a = one
   0 109   3   2   0   0 |   b = at least one
   1  16   7   1   0   0 |   c = at least three
   0  18   1   6   0   0 |   d = at least two
  18   0   0   0   8   0 |   e = two
   2   0   0   0   0   0 |   f = three
```

Figure 1: The confusion matrix of Naïve Bayes classification.

```
=== Summary ===

Correctly Classified Instances         448               84.5283 %
Incorrectly Classified Instances        82               15.4717 %
Kappa statistic                          0.7079
Mean absolute error                      0.0765
Root mean squared error                  0.1952
Relative absolute error                 42.2377 %
Root relative squared error             65.0527 %
Total Number of Instances              530
```

Figure 2: The summary of Naïve Bayes classification.

B. K-nearest Neighbor

Nearest neighbor classification is most often used for classification, although it can be used for other data mining approaches, such as estimation and prediction. It is mainly used when all attribute values are continuous, but it can be modified to deal with categorical attributes. Nearest neighbor estimates the classification of an unseen instance using classification of the instances that are closest to it, in a space that need to be defined. In this method, the attributes are called *dimensions*, and they can be diagrammatically shown if they are small in number.

The distances are usually measured using Pythagora's theorem. If there are two instances in an n-dimensional space, distance between these two instances will be determined using the formula:

$$\sqrt[2]{(a_1 - b_1)^2 + (a_2 - b_1)^2 + \cdots + (a_n - b_n)^2} \quad (5)$$

For Nearest Neighbor approach with categorized values we had choice out of five algorithms (IB1, IBk, KStar, LBR and LWL). We achieved best results with IBk algorithm, with 11 predictor fields (number of times course is attended and grade achieved for every course preceding BP in the chain - OP, OVP, AK and SPA; the code of the group; and part or full time student). Results (the confusion matrix and summary) are shown on Figure 3 and Figure 4.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 338   0   0   0   0   0 |   a = one
   5 109   0   0   0   0 |   b = at least one
   0   0  25   0   0   0 |   c = at least three
   2   0   0  23   0   0 |   d = at least two
   5   0   0   0  21   0 |   e = two
   0   0   0   0   0   2 |   f = three
```

Figure 3: The confusion matrix of Nearest Neighbor classification.

```
=== Summary ===

Correctly Classified Instances         518               97.7358 %
Incorrectly Classified Instances        12                2.2642 %
Kappa statistic                          0.9571
Mean absolute error                      0.0109
Root mean squared error                  0.0641
Relative absolute error                  6.0203 %
Root relative squared error             21.3469 %
Total Number of Instances              530
```

Figure 4: The summary of Nearest Neighbor classification.

C. Decision Tree

One of the most popular classification methods is the construction of a *decision tree*, which contains *decision nodes*, connected by *branches*. These branches are extending from the *root node*, and they terminate with *leaf nodes*. The root node, by convention, is placed at the top of the decision tree diagram, and it describes the attribute which is best suited for initial split according to the decision tree algorithm. Attributes at the decision nodes are tested, and each possible outcome is represented by a branch. Each branch in turn leads either to another decision node, or to a terminating leaf node.

After experimenting with a number of decision trees, namely, DecisionStump, MP5, REPTree, BFTree, ID3, J48, FT, REPTree, AODEsr and UserClassifier, the best results we achieved are with using AODEsr, , with 11 predictor fields (number of times course is attended and grade achieved for every course preceding BP in the chain - OP, OVP, AK and SPA; the code of the group; and part or full time student). Results (the confusion matrix and summary) are shown on Figure 5 and Figure 6.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 337   0   0   0   1   0 |   a = one
   0 111   3   0   0   0 |   b = at least one
   1   9  15   0   0   0 |   c = at least three
   0  10   4  11   0   0 |   d = at least two
  16   0   0   0  10   0 |   e = two
   1   0   0   0   0   1 |   f = three
```

Figure 5: The confusion matrix of Decision Trees classification.

```
=== Summary ===

Correctly Classified Instances         485               91.5094 %
Incorrectly Classified Instances        45                8.4906 %
Kappa statistic                          0.8355
Mean absolute error                      0.0506
Root mean squared error                  0.1454
Relative absolute error                 27.919  %
Root relative squared error             48.4628 %
Total Number of Instances              530
```

Figure 6: The summary of Decision Trees classification.

IV. COMPARISON OF RESULTS

We achieved the best results by using Nearest Neighbor algorithm – 97.7% correctly classified instances and the root mean absolute error is 0.0641, as compared to 82.93% correctly classified instances and 0.107 absolute error in experiments conducted in 2010. As shown in figure 3 most of the elements of the confusion matrix are places at the diagonal of the matrix. The comparison of the results gain in 2010 and 2011 is given in Table 2.

Table 2: Results comparison

Algorithm	Naive Bayes	Nearest Neighbors	Decision Trees
Correctly Classified Instances (2010)	80%	82,93%	83,52%
Correctly Classified Instances (2011)	84,5%	97,7%	91,5%

We can see that there are improvements in all of the used algorithms. Admitting that this improvement is partially due to the 25% larger dataset, we can also not ignore the fact that the results' improvement is achieved primarily through our improved class model.

V. CONCLUSION

Classification is one of the most common methods of data mining, and it helps putting results into "brackets" we use to call classes. In this paper we showed how we can use classification to help us predict student behavior when attending BP course and we have improved prediction about student success at a course using student's previous experience with related courses since 2010. [2] We can see that classification used on our database of students can produce fairly accurate prediction about student success at a course using student's previous experience with related courses. With use of this classification we can have more real expectation from the students just signing on course. What we hope to achieve is to give teachers "heads-up" about what can they expect from the students just signing on their course. We hope that teachers can use our work to "look into" their new student course and decide if students need more basics to help them more easily master course material, or they can start with more advanced topics. Also this research can help to advise students about course selection. Of course, we used only a small fraction of the "Upisi" database, to help us go through with our experiments, but there are a lot of other relationships between different courses that can be exploited and used in betterment of the teaching process.

When we started in 2010 the dataset we used was quite small. In 2011 we used a dataset of 530 instances, cleaned of erroneous instances. We are also aware of the fact that the dataset we can use for now is fairly small, and therefore cannot provide enough insight into the student database, but in one year we achieved much better results. Our next goal is to validate our results (or even improve) in 2012.

REFERENCES

[1] Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons.

[2] Kiroski, Kiril, and Kostoska, Magdalena: Using Classification on Upisi Database, CIIT 2010 Proceedings, Skopje, 2010.

[3] The Gartner Group, www.gartner.com.

[4] David Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.

[5] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 1998.

[6] Discovering knowledge in data: An Introduction to Data Mining, Daniel T. Larose, John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.

[7] Statsoft Electronic Statistics Textbook, http://www.statsoft.com/textbook/data-mining-techniques.

[8] Principles of data mining, Max Bramer, Springer-Verlag London Limited, 2007.

[9] Weka Software, The University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/.