Classifying Multiple Imbalanced Attributes in Relational Data

Amal S. Ghanem, Svetha Venkatesh, Geoff West

Curtin University of Technology, GPO Box U1987, Perth 6845, Western Australia

Abstract. Real-world data are often stored as relational database systems with different numbers of significant attributes. Unfortunately, most classification techniques are proposed for learning from balanced non-relational data and mainly for classifying one single attribute. In this paper, we propose an approach for learning from relational data with the specific goal of classifying multiple imbalanced attributes. In our approach, we extend a relational modelling technique (PRMs-IM) designed for imbalanced relational learning to deal with multiple imbalanced attributes classification. We address the problem of classifying multiple imbalanced attributes by enriching the PRMs-IM with the "Bagging" classification ensemble. We evaluate our approach on real-world imbalanced student relational data and demonstrate its effectiveness in predicting students performance.

1 Introduction

Classification is a critical task in many real-world systems, and is a research field in which extensive studies and experiments are conducted to improve the classification results. A wide range of classification techniques, such as Bayesian networks (BNs), decision trees and Support Vector Machines (SVMs), have been successfully employed in many applications to classify various types of objects.

However, most of these classification techniques are usually proposed with specific assumptions, which may not hold in many real-world domains. The classification of a single attribute from flat data files that have balanced data distribution, represent one of these assumptions. However, in many applications, the collected data are stored in relational database systems with highly imbalanced data distribution, where one class of data has a large number of samples as compared with the other classes. Moreover, in many applications, it is often of interest to classify/predict several attributes rather than a single attribute. An example of such a situation is learning from a student relational database to predict the unit results of a second-year undergraduate student given the results of the first-year, in which the unit results are greatly imbalanced.

Studies have shown that learning from imbalanced data usually hinders the performance of the traditional learning techniques [1,2]. This performance degradation is a result of producing more and stronger rules to classify the samples of the majority class in comparison to that of the minority class, and hence incorrectly classify most of the minority samples to be of the majority class.

Several methods have been proposed to handle the general imbalanced class problem [3–6], and a few attempts have been made to handle the problem particularly in relational data [7–10]. PRMs-IM [10] has been recently introduced as an extension of a relational learning technique: Probabilistic Relational Models (PRMs) [11, 12], to handle the two-class classification of a single imbalanced attribute in relational domains. The idea behind PRMs-IM is to build an ensemble of PRMs on balanced subsets from the original data, in which each subset has an equal number of the minority and majority samples of the imbalanced attribute.

Although the imbalanced class problem is relatively well investigated in both relational and non-relational domains, the classification of several imbalanced attributes in relational domains has not been well addressed. Attempts have been proposed for the special case of classifying two attributes [13,14]. However, these methods did not tackle the imbalanced problem or the relational learning for classifying several attributes.

Therefore, special classification techniques are required to handle the problem of classifying multiple imbalanced attributes in relational domains. In this paper we investigate this problem and review the different proposed approaches. Based on this research, we present a new approach (PRMs-IM2) to handle the problem of classifying multiple imbalanced attributes in relational data. In our approach, we address this problem by combining the balancing concept of PRMs-IM with the "Bagging" classification ensemble [15]. PRMs-IM2 is presented as a Bagging ensemble approach that consists of a set of independent classifiers trained on balanced subsets of the imbalanced data. The subsets are generated using the balancing concept of PRMs-IM for each of the imbalanced attributes. We evaluate our approach on a student relational database with multiple imbalanced attributes, and show the effectiveness of our approach in predicting student results in second semester units.

This paper is organized as follows: section 2 presents a review of the related work. Our methodology is presented in section 3, followed by the experimental results in section 4. Finally, section 5 concludes the paper.

2 Related Work

2.1 Imbalanced class problem in Relational Data

Classification techniques such as BNs, decisions trees and SVMs have been shown to perform extremely well in several applications of different domains. However, several research papers have shown that the performance of these techniques is hindered when applied to imbalanced data [1, 2], as they get biased to the majority class and hence misclassify most of the minority samples.

Methods proposed to handle the imbalanced class problem can be categorized into three groups [16, 17]:

- **Re-sampling:** by either down-sampling the majority class or/and over-sampling the minority class until the two classes have approximately equal

number of samples. A study of a number of down- and over-sampling methods and their performances is presented by Batista et al. [3].
- **Cost-Sensitive learning:** by assigning a distinct misclassification cost for each class, and particularly increasing that of the minority class [4].
- **Insensitive learning:** by modifying the learning algorithm internally to pay more attention to minority class data, as in building a goal oriented BN [5] and exploring the optimum intervals for the majority and minority classes [6].

However, most of these methods are mainly developed for flat datasets, where all data must be presented in one single file. Therefore, in order to learn from a rich relational database, the data must be first converted into a single file that consists of a fixed set of attributes and the corresponding values. This conversion could result in redundant data and inconsistency. Techniques have been proposed to handle the imbalanced class problem in multi-relational data, including: implementing cost-sensitive learning in structured data [7], combining the classification of multiple flat views of the database [8] and using G-mean in decision trees [9].

In addition to these methods, PRMs-IM [10] has been recently introduced to handle the imbalanced class problem in relational data. PRMs-IM was introduced as an extension of the relational learning algorithm: Probabilistic Relational Models (PRMs) [11, 12]. PRMs were introduced as an extension of Bayesian Networks (BNs) to satisfy relational learning and inference. PRMs specify a model for probability distribution over the relational domains. The model includes the relational representation of the domain and the probabilistic schema describing the dependencies in the domain. The PRM model learned from the relational data provides a statistical model that can be used to answer many interesting inference queries about any aspect of the domain given the current status and relationships in the database.

Therefore, to handle the imbalanced class problem in relational data, PRMs-IM was presented as an ensemble of independent PRM models built on balanced subsets extracted from the imbalanced training dataset. Each subset is constructed to include all the samples of the minority class and an equal number of randomly selected samples from the majority class. The number of balanced subsets depends on the statistical distribution of the data. Thus, if the number of samples in the majority class is double that of the minority, then two subsets are created. The PRM models of PRMs-IM are then combined using the weighted voting strategy [10], and hence new samples are assigned to the class with the highest weighted score.

2.2 Classifying Multiple Attributes

Most existing pattern classification techniques handle the classification of a single attribute. However, in many real-world applications, it is often the case of being interested in classifying more than one attribute, such as classifying both

the activity and location in location-based activity recognition systems. The basic solutions for classifying multiple attributes ($\mathcal{A} = \{A_1, A_2, A_3, \ldots\}$) can be classified as follows [13, 14]:

- The combined method: by considering $\mathcal{A}$ as one complex attribute and hence construct one classifier. In this method, advanced techniques are required to work with the multi-class classification.
- The hierarchal method: in a similar approach to decision trees, by constructing a classifier for a given attributes A_i, and then for each class of A_i construct a specialized classifier for A_j. The performance of this method is hindered by the accuracy of the classifiers at the top of the hierarchy, as any misclassification by the top classifiers can not be corrected later. Moreover, in this method, the top attributes can help to reach conclusions about the lower attributes but not vise versa. In addition, the structure grows rapidly as the number of attributes and classes increases.
- The independent method: for each attribute A_i, construct an independent classifier. This method is based on dealing with each attribute separately, and hence it requires more training and testing phases than the other methods.

In addition to these naïve solutions, other methods were proposed but mostly for the special case of classifying two attributes. One method includes using a bilinear model for solving two-factor tasks [14]. This approach mostly acts as a regression analysis and hence does not provide a graphical modelling for interpreting the interactions between the attributes in the domain as provided in other classification techniques, such as BNs.

Another method uses the mutual suggestions between a pair of classifiers [13], in which a single classifier is trained for each attribute, and then at the inference phase, the results of each classifier are used as a hint to reach a conclusion in the other classifier. The learning in this approach is similar to that of the independent approach, but differs in obtaining the final classification results in the inference phase, where the hints between the classifiers are used to reach a better result.

3 Methodology

In this paper we aim to develop a classification technique that could handle the problem of classifying multiple imbalanced attributes in relational data by using the concepts of PRMs-IM [10] and the "Bagging" ensemble approach [15]. PRMs-IM are designed specifically to learn from imbalanced relational databases for a single imbalanced attribute. Thus, for classifying N imbalanced attributes, N independent PRMs-IM models must be performed, one model for each attribute. In PRMs-IM2 we aim to extend PRMs-IM to classify the N imbalanced attributes in a single model.

In order to to obtain a single model, we use the idea of the "Bagging" ensemble approach. The Bagging approach uses an ensemble of K classifiers. Each classifier is trained on a different subset randomly sampled, with replacements,

from the original data. To classify a new sample, the classification decisions of the K classifiers are combined to reach a final conclusion about the class of the sample. A simple combination technique is to use majority voting, in which the sample is assigned to the class with the largest number of votes.

Our approach relies on the idea of building an ensemble of classifiers, where each classifier is trained on a different relational subset that includes a balanced representation of all the imbalanced attributes. This aim is achieved in PRMs-IM2 by firstly applying the balancing concept of PRMs-IM to build balanced relational subsets for each imbalanced attribute. This results in a separate set of balanced subsets for each imbalanced attribute.

However, to achieve the goal of generating subsets that include all the imbalanced attributes, PRMs-IM2 employs the Bagging concept to further sample the balanced subsets into L datasets. Each of the L datasets is formed by randomly selecting one balanced subset form each imbalanced attribute. At the end of this procedure, L balanced subsets will be generated, each subset includes balanced data for each of the imbalanced target attributes. Note that in this paper, we use the same notations to describe the imbalanced situation as those used in [10].

To illustrate our approach, consider a relational dataset $\mathcal{S}$ that consists of a set of attributes $(X_1, X_2, ..., X_M, Y_1, Y_2, .., Y_N)$ organized into tables and relationships, where $(Y_1, Y_2, .., Y_N)$ represents the set of the domain imbalanced attributes that we want to classify. Each Y_i has a majority class $Y_{i_{mj}}$ and a minority class $Y_{i_{mr}}$. In addition, $n_{i(mr)}$ represents the number of samples of the minority class of Y_i. The subsets of PRMs-IM2 are constructed as follows:

- For each imbalanced attribute Y_i of the N imbalanced attributes:
  - Compute n_i as the difference between the number of samples of $Y_{i_{mj}}$ and that of $Y_{i_{mr}}$, where n_i is the number of balanced subsets required for Y_i.
  - For each of the n_i iterations, construct a subset $Y_i s_i$, such that it includes:
    - All the $n_{i(mr)}$ samples from $Y_{i_{mr}}$.
    - $n_{i(mr)}$ randomly selected samples with replacements from $Y_{i_{mj}}$.
    - The data of $(X_1, X_2, ..., X_M)$ according to the selected records of Y_i.
- Compute $L = \max_{i=1..N}(n_i)$, where L is the number of datasets required for the bagging approach.
- For L iterations:
  - Construct S_i database that has the same structure as $\mathcal{S}$.
  - For each Y_j, randomly allocate a subset $Y_j s_k$ from Y_j subsets to S_i.

It is important to note, that when creating the balanced subsets of an imbalanced attribute Y_i, the subsets should include only the data of $(X_1, X_2, ..., X_M)$ and of Y_i. In other words, the data of the other imbalanced attributes $\{(Y_1, Y_2, ..., Y_N)/Y_i\}$ are excluded. This is necessary for creating balanced K databases of all the attributes at the sampling phase. Otherwise, consider the case if Y_i subsets include the related records of Y_j. Then, at the sampling phase, a subset S_k could be generated, such that it includes the random subsets: $Y_i s_l$ and $Y_j s_h$ from Y_i and Y_j, respectively. In this case, the data records of Y_j in S_k will include data from $Y_j s_h$, which are balanced data, and the Y_j records from $Y_i s_l$, which are not balanced.

Having the balanced L relational subsets, an independent PRM model can be learned from each relational subset L_i using the learning techniques described in [12]. Then, these models are combined using the weighted voting strategy as in [10]. In this combination strategy, each PRM model P_i has a different weight $P_{i_w}Y_i$ for each attribute Y_i to be used for the final prediction. The $P_{i_w}Y_i$ is calculated as the average performance accuracy of P_i for classifying Y_i over the training subsets other than the data subset corresponding to P_i. Fig. 1 illustrates the concept of PRMs-IM2.

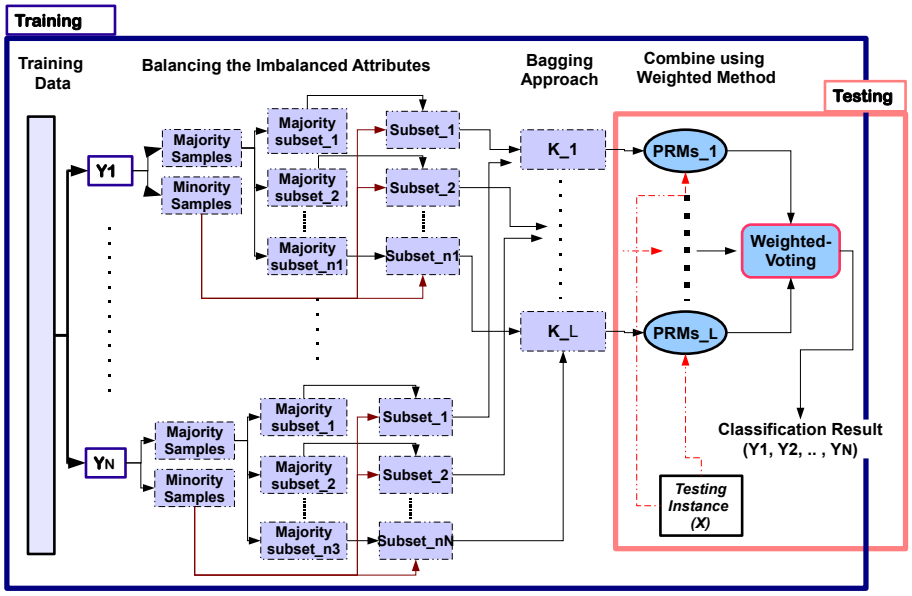

Fig. 1. An illustration of PRMs-IM2 approach for classifying multiple imbalanced attributes.

For a new testing sample x, each P_i outputs the probabilities of each of the imbalanced attributes $(Y_1, Y_2, .., Y_N)$ given the values of $(X_1, X_2, ..., X_M)$. Thus, for each Y_i, each P_i outputs the probabilities $(P_i(x)Y_{i_{mj}}, P_i(x)Y_{i_{mr}})$ for assigning x to $Y_{i_{mj}}$ and $Y_{i_{mr}}$, respectively. Then, for each Y_i, the score of each class equals the summation of the weighted probabilities of the PRM models and hence x is classified to be of the class with the largest weighted score. For example, for Y_i the classification of x is calculated as:

$$F(x) = \text{argmax}_{m \in (Y_{i_{mj}}, Y_{i_{mr}})} (\sum_{\forall P_i} P_i(x)Y_{i_m} * P_{i_w}Y_i) \quad (1)$$

4 Experiments

4.1 Dataset

We use the same Curtin University Student database as used in PRMs-IM [10]. This dataset represents the set of undergraduate students of the Bachelor of Computer Science (BCS) and the Bachelor of Commerce (BCom). The curriculum of the BCS includes: first semester units {ST151, Maths101, FCS151, and English101} and second semester units {ST152 (prerequisite:ST151), FCS152 (prerequisite:FCS151) and IPE151}. The curriculum of the BCom includes: first semester units {BIS100, ACCT100, LFW100 and ECON100} and second semester {MGT100 and MKT100}.

The database includes a set of tables and relationships representing the students' personal information and their performances in first and second semesters of first year. The database is organized as follows:

- The *Personal_Info* table: which consists of: *age, gender, is_international,* and *is_English_home_language* attributes, which each takes values of: {16-19, 20-29, 30-40}, {Male, Female}, {Yes, No}, {Yes, No}, respectively.
- The *Academic_Info* table: which includes: *Preference_no* that takes values of: {1, 2, 3, 4}, to indicate the student's preference of study.
- Semester I units tables: each includes: *grade* of values: {F, 5, 6, 7, 8, 9} representing the categories: {0-49, 50-59, 60-69, 70-79, 80-89, 90-100}.
- Semester II units tables: including the *status* attribute taking values of {Pass, Fail}.

In this dataset, for each of the BCom and BCS degrees, we are interested in predicting a given student's performance in second second semester units based on the personal information and performances in first semester units. However, each of the second semester units represents an imbalanced attribute, in which the majority of data belongs to the majority 'Pass' class compared to few samples belonging to the minority 'Fail' class. Table 1 depicts the data distribution of the training data. For each degree, we perform 5-fold cross validation using the training data for the students enrolled in the period 1999-2005. In addition to the cross validation, we use the data of year 2006 as a separate testing set.

Table 1. Data distribution of (a) the BCS (b) the BCom training dataset.

Unit	$\|Fail\|$	$\|Pass\|$
ST152	12	58
FCS152	11	59
IPE151	7	63

(a)

Unit	$\|Fail\|$	$\|Pass\|$
MGT100	159	1556
MKT100	88	1627

(b)

4.2 Experimental Setup

The results of PRMs-IM2 are presented in comparison to the independent and hierarchal approaches discussed earlier in section 2. In this paper, the combined approach is not evaluated, as it represents a multi-class problem, in which special multi-class algorithms are required. PRM is used as the learning technique in all the experiments as a result of the relational format of the dataset and the effectiveness of PRMs in relational learning.

For evaluation, we use Receiver Operating Characteristics (ROC) curves and the Area under ROC (AUC) [18], which are often used as a measure metric for imbalanced classification problems. ROC curves shows the trade off between the false positive rate and the true positive rate. AUC is used to compare several models using the ROC curves, to get a single value of the classifier performance. The closer the AUC value is to the value '1', the better the classifier.

The independent method is represented as the results of PRMs-IM, in which each independent experiment is evaluated for each imbalanced attribute. In the hierarchal method, the imbalanced attributes are first ordered in descending order $(Y_1, Y_2, .., Y_n)$ based on the AUC value of each attribute obtained in PRMs-IM. Thus, the attributes with higher AUCs are listed first. This order is chosen in order to have the most accurate classifiers at the top of the hierarchy to minimize propagating the classification errors to the lower levels. Moreover, to avoid the problem of the imbalanced class problem, each classifier in the hierarchy is build as a PRMs-IM, thus the classifier of Y_i is a PRMs-IM on balanced subsets of Y_i.

4.3 Experimental results

In this section we present the results obtained from each experiment in terms of: the prediction accuracy in the AUC results and the number of models used for training and inference in each algorithm, as shown in Tables 2 and 3, respectively. For each dataset, the best result is shown in bold. Table 3 presents the normalized number of models of each algorithm for training and inference. The normalized number is the number of models required by each algorithm for a particular dataset divided by the corresponding number of models of PRMs-IM2. Average normalized values greater than one correspond to an algorithm requiring more models than PRMs-IM2.

Table 2. The AUC results (a) Cross validation (b) 2006 testing data

Method	BCom		BCS		
	MGT	MKT	ST	FCS	IPE
PRMs-IM	0.914	0.786	0.839	0.901	**0.913**
PRMs-IM2	**0.922**	**0.893**	**0.950**	**0.923**	0.892
Hierarchal	0.914	0.756	0.811	0.897	**0.913**

(a)

Method	BCom		BCS		
	MGT	MKT	ST	FCS	IPE
PRMs-IM	**0.921**	0.788	0.875	0.927	0.954
PRMs-IM2	**0.921**	**0.840**	**0.984**	**0.968**	**0.993**
Hierarchal	**0.921**	0.787	0.785	0.887	0.954

(b)

Table 3. Normalized number of models used for (a) Training (b) Inference

Method	Dataset (DS) BCom	BCS	Average over DS
PRMs-IM	1.53	2.11	1.82
PRMs-IM2	1.00	1.00	1.00
Hierarchal	2.32	5.56	3.94

(a)

Method	Dataset (DS) BCom	BCS	Average over DS
PRMs-IM	1.53	2.11	1.82
PRMs-IM2	1.00	1.00	1.00
Hierarchal	2.11	3.67	2.89

(b)

In terms of the prediction accuracy, the results show that PRMs-IM2 was able to outperform all the other methods except for the *IPE* dataset in the cross validation. In the hierarchal approach, the results are hindered by the misclassification results of the top classifiers in the hierarchy. In the independent method, the models are built independently for each imbalanced attribute and hence the value of one attribute cannot be used to reach any conclusion about the others. However, in real-world applications, usually the information about one attribute can help to reach better understanding about others. Therefore, a model that includes all the attributes can show the different interactions between them to reach better results. This principle could not be achieved using the independent model, as each attribute needs to be modeled separately, and neither can be accomplished in the hierarchical method, as only the top attributes help to reach a conclusion about the lower attributes but not the other way around. Moreover, the combined approach will treat the targeted attributes as one single attribute and thus would not show us the interactions of each attribute by itself.

This interaction is achieved in PRMs-IM2, as the final model includes all the attributes and presents all the interactions in the domain. Therefore, PRMs-IM2 offers the opportunity for the imbalanced attributes to be related to each other, and hence the value of one of the imbalanced attributes could strengthen the conclusion of the others. Moreover, PRMs-IM2 could model all the significant imbalanced attributes at once and show the different interactions between the attributes, which can not be achieved by the mutual suggestions approach [13] that learns a separate classifier for each imbalanced attribute, or the bilinear model [14] that uses a linear model.

In terms of the number of models used for training and inference, the results show that PRMs-IM2 requires the least number of models for both training and inference. For example in training, the number of models for PRMs-IM and the hierarchy are about twice and four times, respectively, the models of PRMs-IM2, and in inference the number of models are about twice and triple, respectively, those of PRMs-IM.

5 Conclusion

In this paper, we have discussed the problem of classifying multiple imbalanced attributes in relational domains and propose a technique (PRMs-IM2) to handle

this problem. PRMs-IM2 combines the concepts of the relational imbalanced technique (PRMs-IM) and the Bagging ensemble approach. In PRMs-IM2, all the significant imbalanced attributes are modelled in one single model showing the different interactions between the attributes, which can not be achieved by other methods. PRMs-IM2 was evaluated on a student relational database to classify the results of different imbalanced units in second semester. The results show that PRMs-IM2 was able to generally improve over the other naïve methods and at the same time requires the least number of models to perform the training and testing.

References

1. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intell. Data Anal. **6**(5) (2002) 429–449
2. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: ICML. (1997) 179–186
3. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. **6**(1) (2004) 20–29
4. Pazzani, M.J., Merz, C.J., Murphy, P.M., Ali, K., Hume, T., Brunk, C.: Reducing misclassification costs. In: ICML. (1994) 217–225
5. Ezawa, K.J., Singh, M., Norton, S.W.: Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management. In: ICML. (1996) 139–147
6. Kubat, M., Holte, R.C., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Mach. Learn. **30**(2-3) (1998) 195–215
7. Sen, P., Getoor, L.: Cost-sensitive learning with conditional markov networks. In: ICML. (2006) 801–808
8. Guo, H., Viktor, H.L.: Mining Imbalanced Classes in Multirelational Classification. In: PKDD/MRDM'07, Warsaw, Poland (2007)
9. Lee, C.I., Tsai, C.J., Wu, T.Q., Yang, W.P.: An approach to mining the multi-relational imbalanced database. Expert Syst. Appl. **34**(4) (2008) 3021–3032
10. Ghanem, A.S., Venkatesh, S., West, G.: Learning in Imbalanced Relational Data. In: Proc. {ICPR} International Conference on Pattern Recognition, IEEE Computer Society (December 2008)
11. Koller, D., Pfeffer, A.: Probabilistic frame-based systems. In: AAAI. (1998) 580–587
12. Friedman, N., Getoor, L., Koller, D., Pfeffer, A.: Learning Probabilistic Relational Models. In: IJCAI. (1999) 1300–1309
13. Hiraoka, K., Mishima, T.: Classification of double attributes via mutual suggestion between a pair of classifiers. ICONIP **4** (Nov. 2002) 1852–1856 vol.4
14. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Comput. **12**(6) (2000) 1247–1283
15. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2) (1996) 123–140
16. Eavis, T., Japkowicz, N.: A Recognition-Based Alternative to Discrimination-Based Multi-layer Perceptrons. In: Canadian Conference on AI. (2000) 280–292
17. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition **36**(3) (2003) 849–851
18. Fawcett, T.: An introduction to ROC analysis. PRL **27**(8) (2006) 861–874