SVM Learning from Imbalanced Data by GA Sampling for Protein Domain Prediction

Shuxue Zou[1], Yanxin Huang[1], Yan Wang[1], Jianxin Wang[2], Chunguang Zhou[1]
1. College of Computer Science and Technology, Jilin University, China, 130012
Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education
2. School of Information Science and Engineering, Central South University, Changsha,China, 410083
Shuxue Zou, sandror@163.com; Chunguang Zhou, cgzhou@jlu.edu.cn

Abstract

The performance of Support Vector Machines (SVM) drops significantly while facing imbalanced datasets, though it has been extensively studied and has shown remarkable success in many applications. Some researchers have pointed out that it is difficult to avoid such decrease when trying to improve the efficient of SVM on imbalanced datasets by modifying the algorithm itself only. Therefore, as the pretreatment of data, sampling is a popular strategy to handle the class imbalance problem since it re-balances the dataset directly. In this paper, we proposed a novel sampling method based on Genetic Algorithms (GA) to rebalance the imbalanced training dataset for SVM. In order to evaluating the final classifiers more impartiality, AUC (Area Under ROC Curve) is employed as the fitness function in GA. The experimental results show that the sampling strategy based on GA outperforms the random sampling method. And our method is prior to individual SVM for imbalanced protein domain boundary prediction. The accuracy of the prediction is about 70% with the AUC value 0.905.

Keywords: SVM, Imbalanced data, GA, Sampling, Protein domain prediction.

1. Introduction

Imbalanced data learning has recently begun to receive considerable attention for the reason of traditional machine learning methods fail to achieve satisfactory results due to the skewed class distribution. Although Support Vector Machines (SVM) has been extensively studied and has shown remarkable success in many application areas ranging from image retrieval to text classification, the performance of SVM drops significantly [1] when facing imbalanced datasets, where the number of negative instances far outnumbers the positive instances. The factors behind such decrease have already been discussed in other paper [2]. And he have pointed out that it is difficult to avoid such decrease when trying to improve the efficient of SVM on imbalanced datasets by modifying the algorithm itself only .Therefore, as the pretreatment of data, sampling is a popular strategy to handle the class imbalance problem since it re-balances the dataset directly. While the samples selection is a typical combinational optimization problem with exponential complexity.

Recent research on protein domain boundary prediction has been mainly based on widely known machine learning techniques such as Artificial Neural Network, SVM. However the prediction accuracy is not proved subject to the imbalance between the core domain and the boundary.

In this paper, we propose a novel sampling method based on Genetic Algorithms (GA) to rebalance the training dataset for SVM.

Given a query sequence, our method starts by searching the protein sequence database and generating a multiple alignment of all significant hits. The columns of the multiple alignments are analyzed using a variety of sources to define scores that reflect the domain-information-content of alignment columns. Information theory based principles are employed to maximize the information content. We realize the boundary positions are far less than core-domain and take the protein domain prediction as imbalanced data learning problem. In the remainder of this paper, the negatives, i.e., the core domains are always taken to be the majority class and the positives, i.e., the boundary positions, are the minority class. Before SVM learning, sampling based on GA work on. In order to evaluating the final classifiers more impartiality, AUC (Area Under ROC Curve) is employed as the fitness function in GA. The experiment compares the under-sample and over-sample technique in GA sampling. The experimental results show that the sampling strategy based on GA outperforms the random sampling

978-0-7695-3398-8/08 $25.00 © 2008 IEEE
DOI 10.1109/ICYCS.2008.72

method. And our method is prior to individual SVM for protein domain boundary prediction. The accuracy of the prediction is about 70% with the AUC value 0.905.

2. Related Work

2.1. Imbalanced Data

Approaches for addressing the imbalanced training data problem can be divided into two main categories: the data processing approach and the algorithmic approach. The data processing approach can be further divided into two methods: under-sample the majority class, and over-sample the minority class.

2.1.1. SVM. Support Vector machines (SVM) are novel statistical learning techniques that can be seen as typical novel methods for training classifiers based on polynomial functions, radial basis functions, neural networks, splines or other functions. Without loss of generality we choose the SVM coupled with the RBF kernel widely used in pattern recognition.

Given a set of labeled instances $X_{train}=\{x_i, y_i\}$ and a kernel function K, SVM finds the optimal α_i for each x_i to maximize the margin γ between the hyperplane and the closest instances to it. The class prediction for a new test instance x is made through:

$$sign\left(f(x) = -\sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b \right), \quad K(\vec{x}, \vec{x}_i) = \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|^2}{2\sigma^2} \right) \qquad (1)$$

where b is the threshold. The 1-norm soft-margin SVM is used to minimize the primal Lagrangian:

$$L_p = \frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(w \cdot x_i + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n} r_i \xi_i \qquad (2)$$

where $\alpha_i \geq 0$ and $\gamma_i \geq 0$ [2]. The penalty constant C represents the trade-off between the empirical error ξ and the margin. In order to meet the Karush-Kuhn-Tucker (KKT) conditions, the value of α_i must satisfy:

$$0 \leq \alpha_i \leq C \quad and \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (3)$$

2.1.2. SVM Fail to Imbalanced Classification. In [3], Akbani analysis three causes of performance loss with imbalanced data. Firstly positive points lie further from the ideal boundary. And the second is the weakness of Soft-Margins. The last one is imbalanced Support Vector Ratio.

As above, it is difficult to avoid such decrease when trying to improve the efficient of SVM on imbalanced datasets by modifying the algorithm itself only.

2.1.3. Sampling Method. Sampling is a popular strategy to handle the class imbalance problem since it straightforwardly re-balances the data at the data processing stage, and therefore can be employed with any classification algorithm [4]. As one of the successful over-sampling methods, the SMOTE algorithm [5] over-samples the minority class by generating interpolated data. In [6, 7] the integrated sampling technique, with an ensemble of SVM to improve the prediction performance, combines both over-sampling and under-sampling technique.

2.2. Protein Domain Prediction

Protein domains are considered the basic units for protein folding, evolution and function. With the rapid growth in the number of sequences without known structures, it is increasingly important to accurately define protein structural domains and predict domain boundaries. Recent research on protein domain boundary prediction has been mainly based on widely known machine learning techniques such as Artificial Neural Network, SVM. In the various machine learners there are mainly two kinds of information used to feature extraction: protein secondary structure and amino acid composition. DOMpro [8] uses secondary structure, evolutionary information and solvent accessibility information with a recursive neural network; DomSSEA [9] uses predicted secondary structure; SSEPDomain [10] predicts domains by combining information of secondary structure element alignments. On the other hand, Armidillo [11] uses the amino acids composition to predict domain boundaries; the Nagarajan's method [12] is based on analyzing multiple sequence alignments from database searches, position specific physio-chemical properties of amino acids; and DomainDiscovery [13] uses SVM from sequence information including domain linker index. There are also integrated method, such as Albert Y Z. [14] use PSSM, secondary structure, solvent accessibility information and inter-domain linker index to detect possible domain boundaries.

Although a large of number of machine learning based methods have showed their superior performance in protein domain prediction, the overall accuracy of sequence-based methods has been reported in the range of 50 to 70%. The protein domain boundary positions are far less than core-domain. As the analysis in section 2.1, the performance of SVM drops significantly when faced with imbalanced datasets, so as the other traditional machine learners.

3. Proposed Method

3.1. Overview of the Method

Given a query sequence, our algorithm starts by searching the local sequences database and generating a multiple alignment of all significant hits. The columns of the multiple alignments are analyzed using a variety of sources to define scores that reflect the domain-information-content of alignment columns.

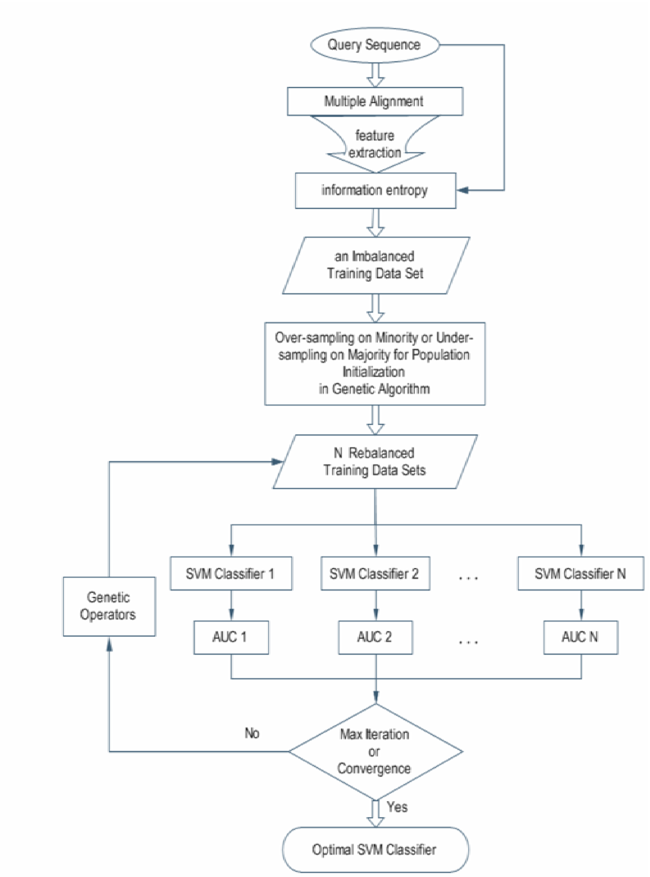
Query Sequence
Multiple Alignment
feature extraction
information entropy
an Imbalanced Training Data Set
Over-sampling on Minority or Under-sampling on Majority for Population Initialization in Genetic Algorithm
N Rebalanced Training Data Sets
SVM Classifier 1
SVM Classifier 2
...
SVM Classifier N
AUC 1
AUC 2
...
AUC N
Genetic Operators
No
Max Iteration or Convergence
Yes
Optimal SVM Classifier

Figure 1 Outline of the proposed method

Information theory based principles are employed to maximize the information content. Besides we get a feature extracted from the conformational entropy of a protein sequence. Thus we get an imbalanced training data set. Next we resample the data set and form N population initialization in Genetic Algorithm. We test respectively the two sampling techniques: over-sampling on minority and under-sampling on majority. SVM learn on each re-sampling training data set and corresponding AUC value is computed. The population is updated by three basic genetic operators, such as reproduction, crossover, mutation, according to the fitness value of AUC. The process of SVM learning and genetic population updated is iterated until convergence or reaching the max iteration. An overview of our method is depicted in Figure 1.

3.2 Datasets and Feature Extraction in Protein Domain Prediction

3.2.1. Datasets. The SCOP database with version 1.65 is employed in this paper, which includes 20,619 proteins and 54,745 chains. The datasets are selected according to the statistical results respectively on the single domain and more than two domains as well as for the consideration of the protein homology.

3.2.2. Feature Extraction. Firstly the query sequence has to been search using BLAST in local protein sequences database. To quantify the likelihood that a sequence position is core domain, or the domain boundary we defined six measures based on the multiple alignments that reflect structural properties of proteins. Information theory based principles are employed to maximize the information content. Besides we quote an approach relating the protein sequence and structure as a domain boundary feature. The simple physical approach is based on the fact that the unique three dimensional structure of protein is a result of the balance between the gain of attractive native interactions and the loss of conformational entropy. All the features detailed computations have been given in our previous paper [15].

3.3. Sampling Based on Genetic Algorithm

One major research direction to overcome the class imbalance problem is to resample the original training dataset, either by over-sampling the minority class or under-sampling the majority class until the classes are represented in a more balanced way. Under-sampling may discard useful data that could be important for the learning process. Over-sampling causes longer training time and inefficiency in terms of memory due to the increased number of training instances and it suffers from high computational costs for preprocessing the data. Thus it is important to select optimal learning samples for the classifier. And the samples selection is a typical combinational optimization problem with exponential complexity.

Genetic algorithms are parallel, global search techniques that emulate natural genetic operators [16]. Because a GA simultaneously evaluates many points in the parameter space, it is more likely to converge to the global solution. Global optimization can be achieved via a number of genetic operators, e.g., reproduction, mutation and crossover. GA is more suitable to the samples selection as GA success in other combinational optimization problem.

A simple Genetic Algorithm is an iterative procedure, which maintains a constant size population N of candidate solutions. During each iteration step (generation) three genetic operators (reproduction, crossover, and mutation) are performing to generate new populations (offspring), and the chromosomes of the new populations are evaluated via the value of the fitness which is related to cost function. Based on these genetic operators and the evaluations, the better new populations of candidate solution are formed.

3.3.1. Genetic Coding and decoding. GA works with a population of binary string. For simplicity and convenience, binary coding is used in this paper. For over-sampling the minority class the duplicated minority index in the original imbalanced training dataset would be coded as binary string of *0*'s and *1*'s with the same length as majority's. In the binary string *1* means the corresponding minority sample would be selected to rebalanced training dataset for the reason of duplication the same sample could be multiply selected. Similarly, for under-sampling the majority class the majority index in the original dataset would be coded as binary string of *0*'s and *1*'s with the same length as minority. In the binary string *0* means the corresponding majority sample would not be selected to rebalanced training dataset. Because of the imbalance between majority and minority, there would be many majority samples lost.

3.3.2. Genetic operators. Crossover is the primary genetic operator, which promotes the exploration of new regions in the search space. For a pair of parents selected from the population the recombination operation divides two strings of bits into segments by setting a crossover point at random, i.e. Single Point Crossover. The segments of bits from the parents behind the crossover point are exchanged with each other to generate their offspring. The mixture is performed by choosing a point of the strings randomly, and switching their segments to the left of this point. The new strings belong to the next generation of possible solutions. The strings to be crossed are selected according to their scores using the roulette wheel. Thus, the strings with larger scores have more chances to be mixed with other strings because all the copies in the roulette have the same probability to be selected.

Mutation is a secondary operator and prevents the premature stopping of the algorithm in a local solution. The mutation operator is defined by a random bit value change in a chosen string with a low probability of such change. The mutation adds a random search character to the genetic algorithm, and it is necessary to avoid that, after some generations, all possible solutions were very similar ones. All strings and bits have the same probability of mutation.

Reproduction is based on the principle of survival of the better fitness. It is an operator that obtains a fixed number of copies of solutions according to their fitness value. If the score increases, then the number of copies increases too. A score value is of associated to a given solution according to its distance of the optimal solution (closer distances to the optimal solution mean higher scores).

3.3.3. Fitness of AUC. The evaluation of a chromosome is done to test its "fitness" as a solution, and is achieved, typically, by making use of a mathematical formula known as an objective function (non-mathematical approaches have also been used). The objective function plays the role of the environment in natural evolution by rating individuals in terms of their fitness. Choosing and formulating an appropriate objective function is crucial to the efficient solution of any given genetic algorithm problem. In our case, selecting the optimal samples set for SVM classifier, a fitness function is the value of Area Under Roc Curve (AUC).

Traditionally, evaluation of a learned model is done by minimizing an estimation of a generalization error or some other related measures [17]. However, accuracy (the rate of correct classification) of a classifier, which is the most frequently used performance measure, is not necessarily a good one. In fact, when the data are strongly imbalanced, accuracy may be misleading since the all-positive or all-negative classifier may achieve a very good classification rate. And situations for which data sets are imbalanced arise frequently in real-world problems and in these cases; model evaluation is done by means of other criteria than accuracy [18].

Metrics extracted from ROC (Receiver Operating Characteristics) curve can be a good alternative for model evaluation, since they can make the difference between errors on positive or negative examples. The most frequently used performance measure extracted from the ROC curve is the value of the area under the curve, commonly denoted as AUC. And Charles [19] proved that AUC is a better measure than accuracy.

The AUC refers to the true distribution of positive and negative instants, but it can be estimated using a sample. The normalized Wicoxon-Mann-Whitney statistic [20] gives the maximum likelihood estimate of the true AUC given *n+* positive and *n-* negative samples:

$$AUC\ (f) = \frac{\sum_{i=1}^{n^+} \sum_{j=1}^{n^-} 1_{f(x_i^+) > f(x_j^-)}}{n^+ n^-} \qquad (4)$$

The two sums in Eq. (4) iterate over all pairs of positive and negative samples. Each pair that satisfies $f(x^+)>f(x^-)$ contributes with $1/(n^+n^-)$ to the overall AUC performance. Thus the AUC can be computed by TPR (true positives / (true positives + false negatives)) and FPR (true negatives / (true negatives + false positives)).

4. Experimental Results

4.1. Parameters Setting

In our experiments, for the SVM, the most important parameters is C and σ^2. According to our previous paper the SVM on the dataset of protein domain prediction is insensitive to the two parameters. Thus we choose the pair of C and σ^2 with (64, 1) for the consideration of the training speed.

For the sampling based on GA, there are several parameters. The number of population in GA is 40 as usual. Two probabilities are involved in genetic operator: the crossover probability is 0.8 and the mutation one is 0.35. If the genetic algorithm could not converge, it will be terminated at maximum iteration=100.

4.2. Speed-up Tricks

During every population updating in GA, the SVM has to be training 2N times, that N is the number of population. Therefore accelerating convergence of GA is the key to the problem. We firstly train the SVM on original imbalanced dataset in order to obtain the support vectors which will be sampled in initial population of GA. On the other hand, for over-sampling on the minority, the SVM will be over learning if there are multiple minorities in initial population. And the GA would not be converged.

4.3. Results

We test respectively the two sampling techniques: over-sampling on minority and under-sampling on majority. The convergence model and experimental results of the two sampling techniques are showed in Figure 2.

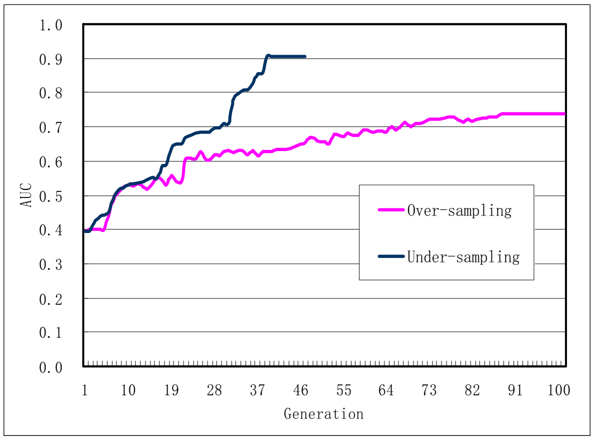

Figure 2 the Comparison of the Over-sampling and Under-sampling Techniques in GA.

The initial AUC is the unique for their identical initial population in GA. While the under-sampling method fast converged to the approximate optimal solution , with the AUC value 0.905, the other is low convergence rate because the search space in over-sampling technique are more than in the under-sampling one. The reason of low AUC value in over-sampling technique is that the SVM over learning leads to the classifier performance degradation.

We also compare our proposed method with others and the results showed in Table 1.

Table 1 the results of different methods

Method		AUC	Accuracy
Individual SVM		0.250	85.74%
SVM with sampling based on GA	Over-sampling	0.715	73.76%
	Under-sampling	0.905	70.2%
SVM with random sampling	Over-sampling	0.472	62.52%
	Under-sampling	0.581	61.89%

Although the individual SVM gain highest accuracy 85.74%, the AUC value is the lowest because the positive samples would be predicted as negative ones because the learned hyperplane of SVM is skewed far from positive side. The duplication of minority in random over-sampling is equal to the corresponding value in over-sampling when the GA converged. But the results of the two methods are different for the reason of the different selection of the minority samples. The experimental results of our method are significantly better than the random sampling method.

5. Conclusion and Future Work

In this paper, we proposed a novel sampling method based on GA to rebalance the imbalanced training dataset for SVM. The samples are coded as binary string represented as selected or not. And AUC is employed as the fitness function in GA to evaluate the performance of SVM classifiers. After implemented the method to the imbalanced problem of protein domain boundary, we got the prediction accuracy of about 70% with the AUC value 0.905. The experimental results show that our method is prior to individual SVM and better than random sampling for the imbalanced training dataset.

Some important issues need to be checked as our future work. Replacing GA with other evolution algorithms in sample selection, such as Particle Swarm Optimization (PSO) and quantum evolutionary algorithm, is the coming works. Designing more suitable genetic operators to accelerate the convergence is another issue. The SVM model complexity strongly depends on the number of surport vectors. When the fraction between support vectors and the training data is close, the sampling method fail and the SVM invalidates for under learning. This need further study.

Acknowledgement

This paper is supported by (1) the National Natural Science Foundation of China under Grant No. 60433020, No. 60673099, No. 6073025 and No. 60773095, (2) National High Technology Research and Development Program of China under Grant No. 2007AAO4Z114, (3) project of science and technology innovation platform of computing and software science (985 engineering), (4) European Commission under Grant No. TH/Asia Link/010 (111084).

6. References

[1] G. Wu, E. Chang, “Class-Boundary Alignment for Imbalanced Dataset Learning”, In ICML 2003 Workshop on Learning from Imbalanced Data Sets II, Washington, DC.

[2] R. Akbani, S. Kwek, N. Japkowicz, “Applying support vector machines to imbalanced datasets”, Proc. 15th. European Conf. Machine Learning (ECML), pp. 39-50, Pisa, Italy, Sep. 2004, 20-24,

[3] Cristianini, N., Shawe-Taylor, J. “An Introduction to Support Vector Machines and other kernel-based learning methods”. Cambridge University Press, Cambridge, 2000.

[4] C. Chen, A. Liaw, L. Breiman, “Using random forest to learn imbalanced data”, Technical Report 666, Statistics Department, University of California at Berkeley 2004

[5] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique”, J. Artif. Intell. Res. (JAIR) 16 ,2002, 321-357.

[6] L. Cen, “Classifying imbalanced data using a bagging ensemble variation (BEV).”, Proceedings of ACM Southeast Regional Conference' 2007. pp.203~208

[7] Yang, L. Aijun, A. Xiangji, H. “Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles”, Lecture Notes in Computer Science, Springer Berlin, 2006.

[8] J. Cheng, M. Sweredoski, P. Bald, “DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks”, Data Mining and Knowledge Discovery 2006, 13(1):1-10.

[9] R.L. Marsden, L.J. McGuffin, D.T. Jones, “Rapid protein domain assignment from amino acid sequence using predicted secondary structure”, Protein Science 2002,11:2814-2824.

[10] J.E. Gewehr, R.Zimmer, “SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles”, Bioinformatics 2006, 22(2):181-187.

[11] M. Dumontier, R. Feldman, H.J. Yao, CWV Hogue, “Armidillo: Domain Boundary Prediction by Amino Acid Composition” , J. Mol Biol 2005, 350:1061-1073.

[12] N. Nagarajan, G. Yona, “Automatic prediction of protein domains from sequence information using a hybrid learning system”, Bioinformatics 2004, 20:1335-1360.

[13] R S. Abdur, Y Z. Albert, “Improving the performance of DomainDiscovery of protein domain boundary assignment using inter-domain linker index”, BMC Bioinformatics. 2006;7 Suppl 1:S6.

[14] P. Yoo, A. Sikder, B. Zhou, A. Zomaya, “Improved general regression network for protein domain boundary prediction”, BMC Bioinformatics. 2008;9 Suppl 1:S12.

[15] Z. Shuxue, H. Yanxin, W. Yan, Z. Chunguang, “A Novel Method for Prediction of Protein Domain Using Distance-Based Maximal Entropy”, Lecture Notes in Computer Science, Springer Berlin, 2007, 1364-1272.

[16] D.E. Goldberg, “Genetic Algorithms in Search Optimization and Machine Learning, Reading”, Addison Wesley, 1989.

[17] V. Vapnik, Statistical Learning Theory, Wiley, 1998.

[18] M. Maloof, “Learning when data sets are imbalanced and when costs are unequal and unknown”, in ICML Workshop on Learning from Imbalanced Data Sets II, 2003.

[19] C. Ling, J. Huang, and H. Zhang, “Auc: a better measure than accuracy in comparing leaning algorithms”, in Proceedings of 2003 Canadian Artificial Intelligence Conference, 2003.

[20] L. Yan, R. Dodier, M. C. Mozer, R. Wolniewicz, “Optimizing classifier performance via the Wilcoxon-Mann-Whitney statistics”. Proceedings of the International Conference on Machine Learning.