Osteoporosis Risk Assessment with Well-Calibrated Probabilistic Outputs

Antonis Lambrou[1,2], Harris Papadopoulos[1,2,3], and Alexander Gammerman[2]

[1] Frederick Research Center, Nicosia, Cyprus
[2] Computer Learning Research Centre, Computer Science Department, Royal Holloway, University of London, England
{A.Lambrou,A.Gammerman}@cs.rhul.ac.uk
[3] Computer Science and Engineering Department, Frederick University, Cyprus
H.Papadopoulos@frederick.ac.cy

Abstract. Osteoporosis is a disease of bones that results in an increased risk of bone fracture. The diagnosis of Osteoporosis is usually performed by measuring the Bone Mineral Density (BMD) using Dual-Energy X-ray Absorptiometry (DEXA) scanning. In this work, we introduce the use of Venn Prediction in order to assess the risk of Osteoporosis before a DEXA scan, based on known risk factors. Unlike other probabilistic methods, Venn Predictors can provide well-calibrated probabilistic outputs under the assumption that the data used are identically and independently distributed (i.i.d.). Our contribution is two-fold: Firstly, we have collected real-world data from various clinic centres in Cyprus which based on their locality can be used for analysis of Osteoporosis risk factors specifically for Cypriot patients. To the best of our knowledge, local data in Cyprus for Osteoporosis risk assessment have not been collected before. Secondary, our results demonstrate that our method can provide probabilistic outputs that may be practical and trustful to physicians.

Keywords: well calibrated probabilities, osteoporosis, risk assessment, Venn Predictor, Machine Learning.

1 Introduction

Osteoporosis is a disease of bones that results in an increased risk of bone fracture. The diagnosis of Osteoporosis is usually performed by measuring the Bone Mineral Density (BMD) using Dual-Energy X-ray Absorptiometry (DEXA) scanning. A result of BMD that is lower than 2.5 standard deviations from the average of young healthy adults is defined by the World Health Organisation (WHO) as Osteoporosis [11].

We introduce the use of Venn Prediction in order to predict the risk of Osteoporosis before a DEXA scan, based on known risk factors. Unlike other probabilistic methods, Venn Predictors can provide well-calibrated probabilistic outputs under the assumption that the data used are identically and independently distributed (i.i.d.). We have collected real-world data from various

H. Papadopoulos et al. (Eds.): AIAI 2013, IFIP AICT 412, pp. 432–441, 2013.

clinic centres in Cyprus which based on their locality can be used for analysis of Osteoporosis risk factors specifically for Cypriot patients. To the best of our knowledge, local data in Cyprus for Osteoporosis risk assessment have not been collected before. Moreover, our results show that Venn Predictors (VPs) can provide probabilistic outputs that may be practical and trustful to physicians.

The rest of the paper is structured as follows. In Section 2, we outline related work. In Section 3, we give a detailed explanation of the Osteoporosis data that we have collected, and we also explain how Venn Prediction works. In Section 4, we show our experimental results on the Osteoporosis data and discuss. Finally, in Section 5, we conclude and outline our future work.

2 Related Work

The World Health Organisation (WHO) has defined the disease of Osteoporosis as a Bone Mineral Density (BMD) which is lower than 2.5 standard deviations from the average of young healthy adults. Furthermore, BMD that is 1 standard deviation lower is defined as Osteopenia, which is a precursor to Osteoporosis [11]. DEXA stands for Dual Energy X-ray Absorptiometry, and is a standard test for BMD. DEXA scanners throw an X-ray beam at the lumbar vertebrae and measure the shadow cast by the bones. In Fig. 1 we include a sample image of the lumbar spine of a DEXA scan. Software in the machine estimates the amount of calcium in the bone based on the darkness of the shadow. The result is expressed as a number of grams per square centimeter, which is defined as the Bone Mineral Density (BMD). In Table 1, we show how the BMD is mapped to a t-score value compared against the average of young healthy adults.

The Venn Prediction (VP) framework is based on the Conformal Prediction (CP) framework. CP is a novel technique for obtaining reliable confidence measures. The technique is proposed in [7] and later improved in [22] and [24]. CPs are built using classical machine learning algorithms, called underlying algorithms. CPs complement the predictions of the underlying algorithms with measures of confidence. Many CPs have been built to date, based on various algorithms such as Support Vector Machines [22], k-Nearest Neighbours for classification [21] and for regression [18], Random Forests [4], and Genetic Algorithms [8]. The computational efficiency of CPs has also been greatly improved using Inductive Conformal Prediction (ICP) [12], as demonstrated in applications to Ridge Regression [17], and more recently in applications to Neural Networks [19]. The CP framework has been successfully applied to medical problems, such as evaluation of the risk of stroke [9], breast cancer diagnosis [6], classification of leukaemia subtypes [1], and acute abdominal pain diagnosis [15]. Additionally, CPs have been applied to other problems such as Software Effort Estimation in [16].

Venn Prediction has been introduced in [23] where the interested reader can find a detailed description of the framework. Since then, VPs have been developed based on k-Nearest Neighbours [3], Nearest Centroid [2] and Neural Networks [13,14]. Furthermore, VPs based on SVMs have been developed in [10,27],

Table 1. Young Adult (YA) T-score based on the Bone Mineral Density (BMD) according to the Wolrd Health Organisation (WHO)

BMD	1.44	1.32	1.20	1.08	0.96	0.84	0.72	0.60
YA T-Score	2	1	0	-1	-2	-3	-4	-5

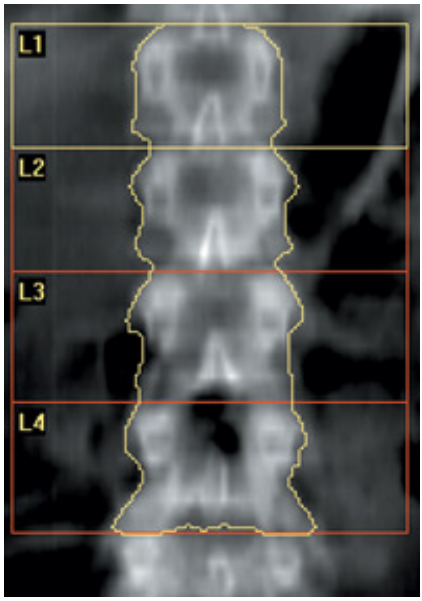

Fig. 1. Image of the Lumbar Spine AP (Anterior Posterior) from a DEXA Scan

and have been compared with Platt's method [20], Binning [5] and Isotonic Regression [26]. As it is shown in [10], such methods do not guarantee that the probabilistic outputs will be well-calibrated.

3 Material and Methods

3.1 Osteoporosis Data

We have collected data from various clinics in Cyprus. In particular, we have collected 389 cases of female patients that have performed a DEXA scan. The data are constructed based on a questionnaire that is given to patients to complete. The patients may have previous history of osteoporosis and may already follow therapy. The questionnaire was constructed by physicians and contains questions that are relevant to Osteoporosis risk factors. Each case is classified as "Normal" or "Risk of Osteoporosis" based on the patient's spine t-score that is given by the DEXA scan. According to the WHO, patients with a t-score above

-1 are diagnosed as healthy, therefore we have classified patients into two classes: "Normal" for patients with t-score above -1, and "Risk of Osteoporosis" otherwise. From the 389 patients, 174 have a t-score above or equal to -1, and 215 have a t-score below -1. In Table 2, we give the list of attributes of our dataset.

3.2 Venn Prediction

In this section, we explain how Venn Prediction (VP) works. Typically, we have a training set[1] of the form $\{z_1, \ldots, z_{n-1}\}$, where each $z_i \in Z$ is a pair (x_i, y_i) consisting of the object x_i and its classification y_i. For a new object x_n, we intend to estimate its probability of belonging to each class $Y_j \in \{Y_1, \ldots, Y_c\}$. The Venn Prediction framework assigns each one of the possible classifications Y_j to x_n and divides all examples $\{(x_1, y_1), \ldots, (x_n, Y_j)\}$ into a number of categories based on a *taxonomy*. A taxonomy is a sequence A_n, $n = 1, \ldots, N$ of finite measurable partitions of the space $Z^{(n)} \times Z$, where $Z^{(n)}$ is the set of all multisets of elements of Z of length n. We will write $A_n(\{z_1, \ldots, z_n\}, z_i)$ for the category of the partition A_n that contains $(\{z_1, \ldots, z_n\}, z_i)$. Every taxonomy $A_1, A_2, \ldots, A_N$ defines a different VP. In this work, we define three VPs based on taxonomies that use the classification output of three underlying algorithms, namely, the J48 decision tree, Random Forests (RF), and Sequential Minimal Optimisation (SMO). Examples are categorized according to the underlying algorithm classifications that are given to them.

After partitioning the examples into categories using a taxonomy, the empirical probability of each classification Y_k in the category τ_{new} that contains (x_n, Y_j) will be

$$p^{Y_j}(Y_k) = \frac{|\{(x^*, y^*) \in \tau_{new} : y^* = Y_k\}|}{|\tau_{new}|}. \tag{1}$$

This is a probability distribution for the class of x_n. So after assigning all possible classifications to x_n we get a set of probability distributions $P_n = \{p^{Y_j} : Y_j \in \{Y_1, \ldots, Y_c\}\}$ that compose the multi-probability prediction of the VP. As proved in [25], these are automatically well calibrated, regardless of the taxonomy used.

The maximum and minimum probabilities obtained for each label Y_k amongst all distributions $\{p^{Y_j} : Y_j \in \{Y_1, \ldots, Y_c\}\}$, define the interval for the probability of the new example belonging to Y_k. We denote these probabilities as $U(Y_k)$ and $L(Y_k)$, respectively. The VP outputs the prediction $\hat{y}_n = Y_{k_{best}}$, where

$$k_{best} = \arg\max_{k=1,\ldots,c} \overline{p(k)}, \tag{2}$$

and $\overline{p(k)}$ is the mean of the probabilities obtained for label Y_k amongst all probability distributions. The probability interval for this prediction is $[L(Y_k), U(Y_k)]$.

[1] The training set is in fact a multiset, as it can contain some examples more than once.

Table 2. Table of attributes in the Osteoporosis dataset

#	Attribute name	Type	#	Attribute name	Type
1	Sex	Binary	35	Receive Thyroxine	Binary
2	Age	Numeric	36	Receive Estrogens	Binary
3	Weight	Numeric	37	Neurogenic Anorexia	Binary
4	Height	Numeric	38	Malabsorption syndrome	Binary
5	Start of Menstruation	Numeric	39	Chronic liver diseases	Binary
6	End of Menstruation	Numeric	40	Inflammatory bowel diseases	Binary
7	Pregnacies	Numeric	41	Transplantation	Binary
8	Smoking now	Binary	42	Chronic renal failure	Binary
9	Smoking in the past	Binary	43	Prolonged immobilization	Binary
10	No smoking	Binary	44	Cushing's syndrome	Binary
11	Years of past smoking	Numeric	45	Epilepsy	Binary
12	Years of current smoking	Numeric	46	Insulin Dependent	Binary
13	Cigarettes per day	Numeric	47	Ovariectomy before menopause	Binary
14	Alchohol intake per day	Numeric	48	Chronic gastrointestinal disorders	Binary
15	Caffeine intake per day	Numeric	49	Paget's Disease	Binary
16	History of fracture	Binary	50	Hyperthyroidism	Binary
17	Hip fracture	Binary	51	Parathyroid gland disease	Binary
18	Spine fracture	Binary	52	Receive Steroids	Binary
19	Wrist fracture	Binary	53	Receive Thyroxine	Binary
20	Low energy	Binary	54	Anticonvulsants (for seizures, epilepsy)	Binary
21	High energy	Binary	55	Diuretics	Binary
22	Sports	Binary	56	Heparin	Binary
23	History of osteoporosis	Binary	57	Chemotherapy	Binary
24	Osteoporosis in family	Binary	58	Treatment of osteoporosis	Binary
25	Loss of height	Binary	59	Alendronati	Binary
26	Kyphosis	Binary	60	Risedronati	Binary
27	End of menstrual bleeding	Binary	61	Zoledronati	Binary
28	Arthritis	Binary	62	Raloxifeni	Binary
29	Secondary Osteporosis	Binary	63	Strontio	Binary
30	Breast feeding	Binary	64	Parathormoni	Binary
31	Avoidance of milk	Binary	65	Denosoymapi	Binary
32	Avoidance of sex	Binary	66	Kalsitonini	Binary
33	Diarrhea	Binary	67	Calcium + Bitamin D	Binary
34	Receive Cortisone	Binary	68	Calcium	Binary

Table 3. Results of the six algorithms on the Osteoporosis dataset. We compare the accuracy, the lower and upper probability outputs of the VPs, and the percentages and accuracy rates of examples which have probabilities of at least 75%, 70%, and 60%.

Predictors	**J48**	**RF**	**SMO**	**J48-VP**	**RF-VP**	**SMO-VP**
Accuracy	70.18%	68.89%	67.10%	67.38%	65.17%	65.71%
Lower Probability	-	-	-	64.27%	57.93%	64.21%
Upper Probability	-	-	-	80.62%	78.09%	71.83%
Min probability $\geq$ 75%	-	-	-	54.73%	34.27%	7.61%
Min probability $\geq$ 70%	-	-	-	67.10%	60.51%	39.49%
Min probability $\geq$ 60%	-	-	-	75.53%	64.88%	84.52%
Accuracy at $\geq$ 75% min. prob.	-	-	-	73.61%	73.43%	74.05%
Accuracy at $\geq$ 70% min. prob.	-	-	-	72.71%	71.51%	69.01%
Accuracy at $\geq$ 60% min. prob.	-	-	-	72.06%	71.44%	67.93%

4 Experiments and Results

4.1 Offline Experiments

We perform 10-fold cross validation experiments on our Osteoporosis dataset with the J48 decision tree, Random Forests (RF), Sequential Minimal Optimisation (SMO), J48-VP, RF-VP, and SMO-VP algorithms. Each algorithm performs a Correlation Based Feature Selection (CBFS) during each fold of the experiment. For the RF, we chose the number of trees to 20, and the depth of each tree to 4 nodes. For the SMO, we chose the RBF kernel with a spread parameter of 0.43. These parameters were chosen after several experimental settings. In the results, we show the average accuracy, and for the VPs we also show the average lower probabilities and upper probabilities. Since VPs provide well-calibrated probabilistic outputs, the accuracy of the VPs is expected to fall within the lower and upper probability intervals. We demonstrate that an amount of predictions have significant probabilities of being correct, and such predictions may be practical for physicians. Moreover, we show that the accuracy rates of the predictions do not drop below the minimum probabilities given by the VPs (up to statistical flactuations).

In Table 3, we compare the average accuracy of the four algorithms on the Osteoporosis dataset. The J48 algorithm provides the highest accuracy which is 70.18%. The corresponding J48-VP provides a slightly lower accuracy which is 67.38%. The VPs provide extra information for each prediction, which is the lower and upper probability interval. We show the average lower probabilities and upper probabilities given by the three VPs. and the corresponding accuracy rates of such predictions. The accuracy rates demonstrate the validity of the probabilistic outputs of the VPs. For example, at 75% probability, the accuracy rates should be at least (up to statistical fluctuations) at the 75% which is expected. In our case, the offline results give accuracies of 73.61%, 73.43%, and 74.05%, which are acceptable. In the table, we also show the percentage of examples that have a minimum probability of 75%, 70%, and 60%. The J48-VP

Table 4. Confusion matrices of the three algorithms J48, RF, and SMO on the Osteoporosis dataset

	J48		**RF**		**SMO**	
	Normal	Risk of Ost.	Normal	Risk of Ost.	Normal	Risk of Ost.
Normal	113	61	107	67	89	85
Risk of Ost.	55	160	54	161	43	172

gives 54.73% of predictions with a probability of at least 75%. These predictions can be considered significant, since the error rate will be at most 25% in the long run.

In Table 4, we show the confusion matrices of the three algorithms on the Osteoporosis dataset. The confusion matrices demonstrate that the algorithms misclassify examples at a balanced rate between the two classes.

4.2 Online Experiments

In order to show the validity of the probability estimates of the VPs, we conduct experiments in the on-line mode. Initially all examples are test examples and they are added to the training set one by one after a prediction for each one is made. We graph the Cumulative Lower Accuracy Probability (CLAP), the Cumulative Upper Accuracy Probability (CUAP), and the Cumulative Accuracy (CA) curves:

$$CLAP(t) = \frac{1}{t}\sum_{i=1}^{t} U_i(Y_{k_{best}}), \tag{3}$$

$$CUAP(t) = \frac{1}{t}\sum_{i=1}^{t} L_i(Y_{k_{best}}), \tag{4}$$

$$CA(t) = \frac{1}{t}\sum_{i=1}^{t} Acc_i, \tag{5}$$

where t is the number of test examples that have been added to the training set, and $Acc_i = 1$ when the prediction for example x_i is correct and 0 otherwise. Since VPs provide well calibrated probabilistic outputs, it is expected that the CA curve will fall within or near the CLAP and CUAP curves.

In Fig. 2, we show the online results of the three VPs, J48-VP (left), RF-VP (right), and SMO-VP (bottom) on the Osteoporosis dataset. The results demonstrate the validity of the VPs, since the actual accuracy of the predictors always falls within the lower and upper probability outputs of the VPs. Comparing the three VPs, we can see that the RF-VP gives the wider probabilistic outputs, while the SMO-VP gives the narrowest outputs.

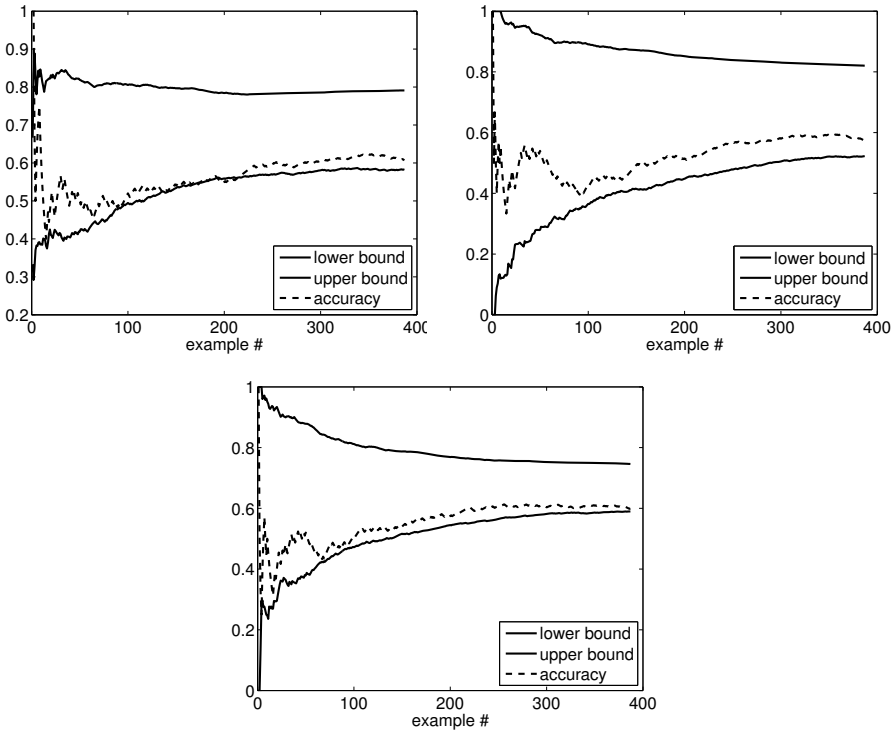

Fig. 2. Online experiments with J48-VP (left), RF-VP (right), and SMO-VP (bottom) on the Osteoporosis dataset

5 Conclusion

In this work, we have applied Venn Prediction to the problem of Osteoporosis Risk Assessment. We have evaluated our method on real-world data that we have collected from various clinics in Cyprus. Our results, demonstrate that our method provides well-calibrated probabilistic outputs in the predictions that can be useful in practice. Precisely, patients may get a prognosis based on Osteoporosis risk factors before the performance of a DEXA scan. In the future, we aim to collect more data and perform supplementary analysis, in order to improve and evaluate further our VPs. Furthermore, we are in the process of building a tool for physicians that will enable them to use our VPs for assessing the risk of Osteoporosis.

Acknowledgments. This work was co-funded by the European Regional Development Fund and the Cyprus Government through the Cyprus Research Promotion Foundation "DESMI 2009-2010", TPE/ORIZO/0609(BIE)/24 ("Development of New Venn Prediction Methods for Osteoporosis Risk Assessment").

References

1. Bellotti, T., Luo, Z., Gammerman, A.: Reliable classification of childhood acute leukaemia from gene expression data using confidence machines. In: Proceedings of IEEE International Conference on Granular Computing (GRC 2006), pp. 148–153 (2006)
2. Dashevskiy, M., Luo, Z.: Reliable probabilistic classification and its application to internet traffic. In: Huang, D.-S., Wunsch II, D.C., Levine, D.S., Jo, K.-H. (eds.) ICIC 2008. LNCS, vol. 5226, pp. 380–388. Springer, Heidelberg (2008)
3. Dashevskiy, M., Luo, Z.: Predictions with confidence in applications. In: Perner, P. (ed.) MLDM 2009. LNCS, vol. 5632, pp. 775–786. Springer, Heidelberg (2009)
4. Devetyarov, D., Nouretdinov, I.: Prediction with Confidence Based on a Random Forest Classifier. In: Papadopoulos, H., Andreou, A.S., Bramer, M. (eds.) AIAI 2010. IFIP AICT, vol. 339, pp. 37–44. Springer, Heidelberg (2010)
5. Drish, J.: Obtaining calibrated probability estimates from Support Vector Machines (1998)
6. Gammerman, A., Nouretdinov, I., Burford, B., Chervonenkis, A., Vovk, V., Luo, Z.: Clinical mass spectrometry proteomic diagnosis by conformal predictors. Statistical Applications in Genetics and Molecular Biology 7(2) (2008)
7. Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. In: Uncertainty in Artificial Intelligence, pp. 148–155. Morgan Kaufmann (1998)
8. Lambrou, A., Papadopoulos, H., Gammerman, A.: Reliable confidence measures for medical diagnosis with evolutionary algorithms. IEEE Transactions on Information Technology in Biomedicine 15(1), 93–99 (2011)
9. Lambrou, A., Papadopoulos, H., Kyriacou, E., Pattichis, C.S., Pattichis, M.S., Gammerman, A., Nicolaides, A.: Evaluation of the risk of stroke with confidence predictions based on ultrasound carotid image analysis. International Journal on Artificial Intelligence Tools 21(04), 1240016 (2012)
10. Lambrou, A., Papadopoulos, H., Nouretdinov, I., Gammerman, A.: Reliable probability estimates based on support vector machines for large multiclass datasets. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H., Karatzas, K., Sioutas, S. (eds.) AIAI 2012 Workshops, Part II. IFIP AICT, vol. 382, pp. 182–191. Springer, Heidelberg (2012)
11. World Health Organisation. Prevention and management of Osteoporosis, Geneva (2003)
12. Papadopoulos, H.: Inductive Conformal Prediction: Theory and application to Neural Networks. In: Fritzsche, P. (ed.) Tools in Artificial Intelligence, ch.18, pp. 315–330. InTech, Vienna (2008)
13. Papadopoulos, H.: Reliable probabilistic prediction for medical decision support. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) EANN/AIAI 2011, Part II. IFIP AICT, vol. 364, pp. 265–274. Springer, Heidelberg (2011)
14. Papadopoulos, H.: Reliable probabilistic classification with neural networks. Neurocomputing 107, 59–68 (2013)
15. Papadopoulos, H., Gammerman, A., Vovk, V.: Reliable diagnosis of acute abdominal pain with conformal prediction. International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications 17(2-3), 127–137 (2009)
16. Papadopoulos, H., Papatheocharous, E., Andreou, A.S.: Reliable confidence intervals for software effort estimation. In: Proceedings of the 2nd Workshop on Artificial Intelligence Techniques in Software Engineering (AISEW 2009). CEUR Workshop Proceedings, vol. 475, pp. 211–220. CEUR WS.org (2009)

17. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, pp. 345–356. Springer, Heidelberg (2002)
18. Papadopoulos, H., Vovk, V., Gammerman, A.: Regression conformal prediction with nearest neighbours. Journal of Artificial Intelligence Research 40, 815–840 (2011)
19. Papadopoulos, H., Vovk, V., Gammerman, A.: Conformal prediction with neural networks. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), vol. 2, pp. 388–395. IEEE Computer Society (2007)
20. Platt, J.C.: Probabilistic outputs for Support Vector Machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74. MIT Press (1999)
21. Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A.: Transductive confidence machines for pattern recognition. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, pp. 381–390. Springer, Heidelberg (2002)
22. Saunders, C., Gammerman, A., Vovk, V.: Transduction with confidence and credibility. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence, vol. 2, pp. 722–726. Morgan Kaufmann, Los Altos (1999)
23. Vovk, V., Shafer, G., Nouretdinov, I.: Self-calibrating probability forecasting. In: Thrun, S., Saul, L.K., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems 16, pp. 1133–1140. MIT Press, Cambridge (2004)
24. Vovk, V., Gammerman, A., Saunders, C.: Machine-learning applications of algorithmic randomness. In: Proceedings of the Sixteenth International Conference on Machine Learning, pp. 444–453. Morgan Kaufmann (1999)
25. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)
26. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining, pp. 694–699 (2002)
27. Zhou, C., Nouretdinov, I., Luo, Z., Adamskiy, D., Randell, L., Coldham, N., Gammerman, A.: A comparison of venn machine with platt's method in probabilistic outputs. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) EANN/AIAI 2011, Part II. IFIP AICT, vol. 364, pp. 483–490. Springer, Heidelberg (2011)