Improving the Decision Support in Diagnostic Systems Using Classifier Probability Calibration

Xiaowei Kortum[1(✉)], Lorenz Grigull[2], Urs Muecke[2], Werner Lechner[3], and Frank Klawonn[1,4]

[1] Department of Computer Science, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, 38302 Wolfenbuettel, Germany
{x.kortum,f.klawonn}@ostfalia.de
[2] Department of Paediatric Haematology and Oncology, Medical University Hanover, Carl-Neuberg Str.1, 30625 Hannover, Germany
{grigull.lorenz,urs.muecke}@mh-hannover.de
[3] Improved Medical Diagnostics IMD GmbH, Ostfeldstr. 25, 30559 Hannover, Germany
werner.lechner@improvedmedicaldiagnostics.com
[4] Helmholtz Center for Infection Research, Inhoffenstrasse 7, 38124 Braunschweig, Germany
Frank.Klawonn@helmholtz-hzi.de

Abstract. In modern medical diagnoses, classifying a patient's disease is often realized with the help of a system-aided symptoms interpreter. Most of these systems rely on supervised learning algorithms, which can statistically extend the doctor's logic capabilities for interpreting and examining symptoms, thus supporting the doctor to find the correct diagnosis. Besides, these algorithms compute classifier scores and class labels that are used to statistically characterize the system's confidence level on a patient's type of disease. Unfortunately, most classifier scores are based on an arbitrary scale but not uniformed, thus the interpretations often lack of clinical significance and evaluation criterion. Especially combining multiple classifier scores within a diagnostic system, it is essential to apply a calibration process to make the different scores comparable.

As a frequently used calibration technique, we adapted isotonic regression for our medical diagnostic support system, to provide a flexible and effective scaling process that consequently calibrates the arbitrary scales of classifiers' scores. In a comparative evaluation, we show that our disease diagnostic system with isotonic regression can actively improve the diagnostic result based on an ensemble of classifiers, also effectively remove outliers from data, thus optimize the decision support system to obtain better diagnostic results.

Keywords: Classifier calibration
Isotonic regression · Pool adjacent violators
Multiple-classifier system · Statistical computing

H. Yin et al. (Eds.): IDEAL 2018, LNCS 11314, pp. 419–428, 2018.
https://doi.org/10.1007/978-3-030-03493-1_44

1 Introduction

Compared with the common disease appearances, diagnosing rare types based on a patient's symptoms is hard to achieve. In particular, when multiple influencing factors need to be taken into account, overlooked or inaccurately interpreted symptoms of a rare disease are common. Even when the preliminary examination for a patient's physical condition and laboratory tests have been completed for delineating the range of possible diseases, potential lack of awareness and data resources make rare diseases still hard to be identified for many medical doctors. De facto, classifying rare diseases involves a sophisticated process which can take years from early symptom appearance to a final diagnosis. Solutions like a system-assisted diagnosis for improved interpretation of signs through collective patient records are a frequently requested strategy by medical institutions.

A collaborative research project between scientific researchers and medical experts from Hannover Medical School (MHH) has been initiated. The original idea of this cooperation is based on realizing a computer-aided diagnostic system that supports medical doctors' classification capabilities on new patients with symptoms for specific types of rare diseases. For effectively collecting patient data, MHH designed questionnaires for particular disease groups through interviews, investigations, and observations of patients that have already been diagnosed. Such questionnaires cover several fields with the focus of significant symptoms and binary classification as doctor mentioned result. The combination of question & answer pattern provides strong evidence for a particular disease.

In previous publishings of our study, we could manifest a fusion classifier diagnostic system [7] that unites and channelizes the qualitative benefits of single classifier, i.e. support vector machines (SVM) [2], linear discriminant analysis (LDA) [12], logistic regression (LR) [12] and random forests (RF) [8]. The core of this system relies on methods that focus on assisting medical diagnostics [9], especially the previous established fusion method takes advantages of each classifier through evaluating their compatibility and accuracy of each derived diagnosis. Experimental approaches revealed that the fusion recognition, as a combination of various classifiers, presents a definite performance improvement rather than any single classifier [13].

Each classifier provides probabilities for the considered diseases. Some of the classifiers predict well-calibrated probabilities because they do not have biases. Some maximum margin algorithms such as SVM tend to push predicted probabilities away from 0 and 1, which brings characteristic sigmoid shaped distortion [3]. Other models such as Naive Bayes rely on unrealistic independence assumptions and tend to push probabilities closer to 0 and 1. Based on different classifiers have inequable properties, it is essential to narrow the gap of classifiers' scales and consider well-calibrated probabilities while ensemble multiple classifiers [11].

Isotonic regression is a powerful calibration method that can help in correcting monotonic distortions and rescale classifier probabilities into the same range, as the example shows in Fig. 1. The calibration module allows us to obtain better comparable probabilities of a given model, therefore enable a more meaningful combination of classifier outputs. Which makes it easier for the decision

support system to rank examples in order of class-membership likelihood and find a more accurate probability of a new patient that belongs to a particular class, thus reach a higher accuracy of decision making [15]. Another reason to do a calibration on classifier scores is that classes are often unbalanced. It is helpful to harmonize the data by introducing bias to underrepresented classes [11].

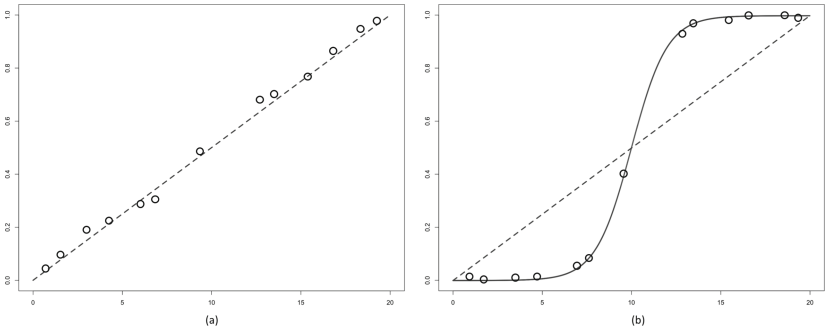

Fig. 1. Original probability (a) vs. Calibrated probability (b)

2 Related Work

This work relies on previous achievements and related studies with particular focus on multiple classifier systems, and best practice for the system-based decision improvements using probability scaling. The following concepts were methodologically adapted for the targeted medical diagnostic environment.

Zadrozny et al. [15] introduced the Pool Adjacent Violators Algorithm (PAVA) that enables the calibration of multi-class probability estimates in applying a simple ranked mapping methodology on classifier scores. The probability is an essential factor which represents the confidence level on the predicted outcome.

Niculescu-Mizil et al. [11] examine the probabilities predicted by ten supervised learning algorithms and the effectiveness of isotonic regression for calibrating the predictions made by different learning methods. It is shown that after calibration, most classifier models can predict better probability estimates.

Kortum et al. [7] investigated the benefits of using the classifier fusion method for diagnosing rare diseases in practice. Experimental results show that this strategy dramatically improves the accuracy of the system compared to any single classifier. An auxiliary tool for physicians was implemented that could derive computer-aided diagnoses in comparing a new patient's symptoms with classified records from a shared patient database.

Chen et al. [3] further investigate the characteristic problem of classifier scores in medical diagnostic approaches. The authors proved that classifier scores on an arbitrary scale could be converted to the probability scale for a target population prevalence value without affecting discrimination performance. This result takes

an essential role in this paper since the dedicated probability problem addresses an interpretability gap, which solves a potential risk in classification.

3 Diagnostic Process for New Patient Data

In this chapter, we describe the classification improvement process for our computer aided diagnosis system that was originally revealed in our previous study stages [6,7]. Figure 2 provides a compact structure overview about the decision architecture when classifying a new patient's symptoms while considering probability calibration methods, e.g. the isotonic regression.

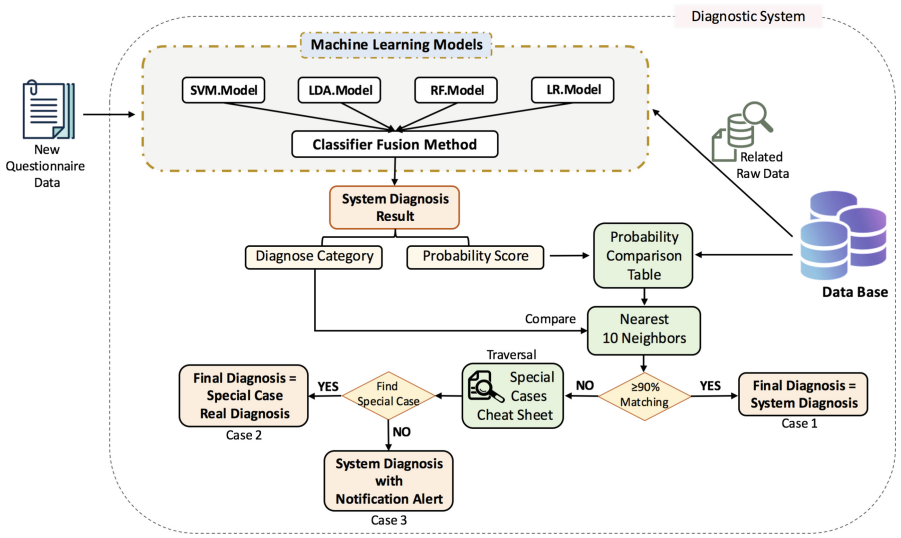

Fig. 2. System diagnosis process for a new patient

When a filled in questionnaire of a new patient is available for diagnosis, the system's first step is to extract the questionnaire related raw data from the database as the foundation of classification models. The classifier fusion method involves four algorithms (Support Vector Machines; Linear Discriminate Analysis; Random Forest and Logistic Regression) as a combination of multiple classifiers with robust prediction capabilities on medical data records in supervised modeling approaches [7]. After training each classification model by raw data, the system proceeds in predicting the probability of new coming data from each classifier. The fusion method is applied to obtain the diagnostic result due to the highest corresponding probability score, denoted as P.

Then the acquired P and its associated classification result will be compared within the range of its closest K-neighbors in the Probability Comparison Table

(described in Sect. 3.3), K depends on the dataset size and the degree of dispersion of data within this interval. The comparison derives how many real doctors' decisions in a similar case are matching the obtained system diagnostic result. If the prediction presents a strong diagnosis probability ($\geq$90% matched), the system decision will be selected as the final result. In case of a weak prediction value (<90%), the new patient symptoms data will be sequentially compared with the special case cheat sheet to check if there are any similar answer patterns in the database records with a firm diagnosis. A decisive matching (Case 2) will lead to the final diagnose according to the real doctors' diagnosis of the matched cases. If it does not find any matches in the cheat sheet (Case 3), a prompt message will be given along with the system diagnosis, indicating which variable affects the system to make the correct diagnosis because of insufficient evidence, thus helps doctors to examine the related symptoms tendentiously. Once the patient' diagnostic has been confirmed, his significant rare data will be collected and backing-up for documentation and discussion purposes. Compared with the initial diagnosis system, the subsequent screening process presents an additional improvement in filtering different cases between strong statistical predictions and poor predictions. In this way, it is clearly indicated how reliable the diagnostic proposal of the system is.

3.1 Classifier Fusion Method

The concept of applying multiple classifiers relies on the idea that the fusion method takes advantages of each classifier's outcome through evaluating their compatibility and accuracy, to derive the optimal computer-assisted diagnosis. Our previous studies proved that classifiers ensemble for symptoms interpretation, especially the combination of different supervised learning algorithms as shown in Fig. 2, present strong performance improvements compared with any single classifier [7,11].

The diagnostic process assigns a new dataset of a patient's disease symptoms to the trained classification patterns. The classifier fusion method applied in this study derives the average score from four supervised classifier scores *P(average)*. Each represents the likelihood for a particular type of disease d within the continuous range from 0 to 1. Further, each system-aided diagnosis d obtains the average probability value *P(d)* as a representative indicator of a particular class. The diagnostic class from the highest *P(average)* becomes selected as the diagnosis outcome. By evaluating the compatibility and accuracy of individual classifiers, the fusion method takes advantages of each single classifier [9]. In case that one classifier is more particular about a specific diagnosis, it will occupy a more significant proportion within the fusion method. This allows a medical practitioner to derive more accurate diagnosis in comparing a new patients symptoms with the exact diagnostic records from the related dataset.

3.2 Isotonic Regression for Calibrating Classifier Scores

The classifier fusion model performs its system-based diagnoses through statistical classifiers, thus the outcome can generally be found within a limited scale. Toward that, the medical diagnostic meaning of classified decision scores is not always easy to interpret [15]. Modeling that involves the concept of calibration provides advantages when transferring derived classifier scores to a more meaningful probability space. The diagnosis of a disease is more reliable when the doctor can directly interpret the prediction score.

Therefore, the unnormalized scores produced from classifiers need to be calibrated to score-conditional probabilities, to supply reliable performance measures for generalization in the medical field [3]. For example, a patient's classified diagnosis score of 0.80 presents a high likely indication for a particular disease – according to the average of all previously experienced patients. However, a patient with a classifier score of 0.45, on the other hand, does not clearly distinguish between types of diseases without straight comparison to other classified patient records. This leads to the concern about the discrimination and calibration of probabilities within multi-classifier systems [5].

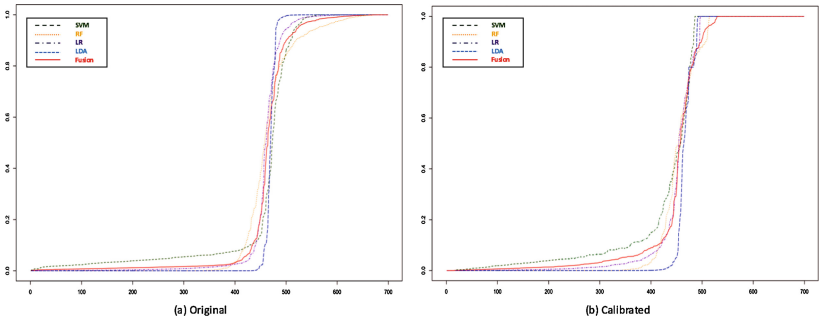

Fig. 3. Original classifier scores vs. calibrated scores - dataset: BC

Fig. 3 shows the difference between original classifier scores (a) and calibrated classifier scores (b), by using the Wisconsin breast cancer dataset (*BC*) from the UCI repository [4]. The calibrated classifier scores (b) tend to interpret the probability more closely as binary scoring without manipulating the models' posterior probabilities. This can be improved with the help of an optimal threshold selection for binning the scores. The initial maximum margin of methods such as SVM push the posterior probability away from 0 and 1 while techniques such as LDA tend to push the probability towards 0 and 1. Methods like RF and SVM perform much better after applying isotonic regression for calibration, which helps to transform classifier scores into meaningful probabilities. Isotonic regression is a non-parametric technique that does not make any assumptions such as linearity among variables and constant error variance [14].

3.3 Integrated Probability Comparison Table

The Probability Comparison Table can be obtained by applying 10-Fold-Cross-Validation to traverse the entire raw data. As shown in Fig. 4, the operation process is divided into three steps: Classifier Model Training; Calibration Model Training and Fuse Prediction Scores.

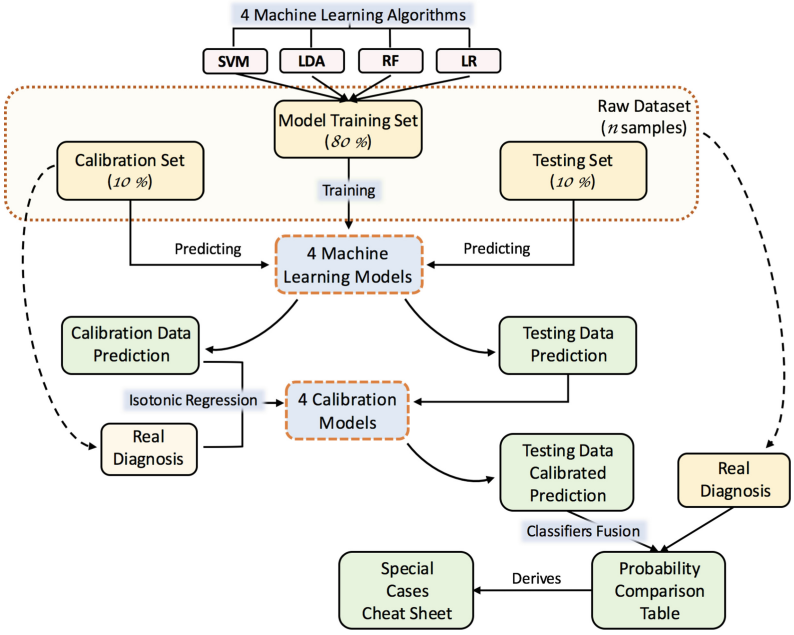

Fig. 4. Fusion classifier calibration for minimized classification errors

In each cross-validation iteration, the applied dataset with one column's binary classification will be partitioned into three parts: Model Training Set (80%); Calibration Set(10%) and Testing Set (10%). As the first step, 80% of the data records are used for training the classification models. The calibration model needs two elements to train: a prediction score and a corresponding real diagnosis. Thence before training the calibration model, the Calibration Set without diagnosis column will go through the model prediction process to obtain the corresponding predictive value. By using the isotonic regression along with the previously detached diagnosis column, the regular prediction values are used to train the calibration models. Probability calibration methods take commonly place for scaling each classification scores' range into a directly-viewed and easy to understand prevalence rate [15]. This step is analogous to the applied testing, the alternative testing data prediction, except that its probability outcome has improved statistical characteristics due to the applied isotonic regression.

As a testing set that separates the doctors' original diagnosis column beside, it needs to be applied to four trained classification models and four calibration models to get calibrated prediction probabilities finally. After applying the fusion method [6,7], the Probability Comparison Table will be generated, for later comparison and examination with the final system diagnosis. The table contains each questionnaire's original classifier scores, corresponding calibrated probability estimation, and real doctors' diagnoses of similar patient cases.

4 Evaluation of Disease Diagnostic System

We applied leave-one-out cross-validation (LOOCV) [1] to evaluate how well the improved model in Fig. 2 performs on new data, by determining the diagnostic accuracy for patients correctly interpreted disease symptoms. The improved system will be compared with the initial system by their qualitative overall accuracy. When it comes to supervised learning algorithms, LOOCV presents a common strategy to quantify the system's decisional accuracy. We evaluated the system with three different datasets:

- Breast cancer dataset (BC) with 699 records [4]
- Primary immunodeficiency disorders dataset (PID) with 126 records [10]
- Rare disease dataset (RD) with 1021 records [6]

All the data records consist of the same characteristic: multiple attribute columns that describe the patients' symptoms condition and a binary diagnosis column given by a real doctor. The data we use in this paper are from patients that have been individually examined, tested and diagnosed by medical doctors.

The procedure of LOOCV assesses our model quality iteratively, by removing one sample from the dataset and use it as testing data. The remaining samples are then used for training the model and predict the diagnosis category of the sample data that have been left out. After all data records have been sequentially extracted and diagnosed by the system, the model's overall accuracy can be calculated by the percentage of correct diagnoses compared with real doctors' decisions. Table 1 shows the system accuracy for the three different datasets. The second column lists the systems overall accuracy. The last column, which is the original diagnostic system accuracy for a classifier fusion without the use of isotonic regression; it can be seen that our strategy (shows in Fig. 2) can significantly enhance the overall accuracy of the decision support system.

The improvement for diagnosing breast cancer increases 1%, foremost because the initial system through the fusion classifier could already make a reliable diagnosis accuracy of 96.1%. The other two datasets achieve a better improvement with 3.6% for the primary immunodeficiency disorders recognition and 3.1% for rare disease diagnosis. Experimental result demonstrates that the designed system can improve the overall accuracy of disease diagnostics, thus provide influential decision support for doctors' diagnosis. In addition to the overall accuracy, another critical information is system verification represented

Table 1. Disease diagnostic system overall accuracy for three datasets

Data set	Improved system accuracy	Case 1 accuracy	Caes 2 accuracy	Case 3 accuracy	Original system accuracy
BC	97.1%	98.6%	100%	92.7%	96.1%
PID	88.9%	95.2%	99.8%	71.7%	85.3%
RD	86.4%	93.5%	99.1%	66.6%	83.3%

in three different cases. Case 1 ensures the high prediction accuracy by comparing the system diagnosis results with the doctors' decision within similar cases, which required to achieve at least 90% matches. In the second case, the diagnosis accuracy could reach 99%, reveals that the diagnostic result of some patients that have a relatively weak matching degree in case 1, could be adjusted by matching their record with particular response patterns in which the system cannot make an accurate decision (records into cheat sheet). Due to the similar answer patterns as in the reference records, the patient's diagnosis result will follow the doctors' suggestions, which are substantiated correct in most case. Mainly, if there is a one-to-one correspondence between a new patient's answer pattern and the special case in the cheat sheet, the system will ensure that the final diagnostic results are consistent with the doctor's decision stored in the database, regardless of the system diagnosis. Case 3 mainly collects data records that the system cannot diagnose accurately. It is possible that all classifiers point to one diagnosis category, but the reality is poles apart. In this case, the best way is to mark the variables that influence the system diagnosis and send notifications to provide constructive suggestions to doctors for examining related symptoms. Once the doctor gets confirmation of patient' diagnostic result, such significant data record will be absorbed into our cheat sheet, for more in-depth analysis and discussion purposes.

5 Conclusion and Future Perspectives

Motivated by continuously improving the disease diagnostic system, we embedded the calibration procedure into the classifiers fusion method. The conventional classification model can only interpret patients' diseases within each machine learning algorithm. But the newly integrated calibration adaption based on isotonic regression offers the possibility of cross-reference classifiers' results in the same interval, thereby derive diagnoses through three degrees of qualified outcomes (firm statistical determination, answer pattern matching and handling significant rare data record). We validated our diagnostic system by using three datasets covering 1846 records that include multiple diseases symptoms description and real doctors' diagnoses. We compared the overall accuracy of the original fusion classifier model and the calibrated classification, proved a new strategy that can improve system performance. The system can support doctors to derive

the correct diagnosis. More importantly, it can filter out the poor diagnostic outcome. We believe that the fusion classifier calibration and its case differentiation could help researchers and medical doctors to improve their patients' symptoms interpretation, also in other data analysis areas. It presents an improved measure for adequate diagnosis, especially for patients with symptom tendencies where the decidability between two types of diseases is not explicitly apparent.

References

1. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Dev. Appl. Stat. **4**(2010). 4079, 159–177 (2003)
2. Auria, L., Moro, R.A.: Support vector machines (SVM) as a technique for solvency analysis (2008)
3. Chen, W., Sahiner, B., Samuelson, F., Pezeshk, A., Petrick, N.: Calibration of medical diagnostic classifier scores to the probability of disease. Stat. Methods Med. Res. **27**(5), 1394–1409 (2018)
4. Wolberg, W.H., Street, W.N., Mangasarian, O.L.: UCI machine learning repository: breast cancer wisconsin (1995). http://archive.ics.uci.edu/ml/datasets
5. Schmid, C.H., Griffith, J.L.: Multivariate Classification Rules: Calibration and Discrimination. American Cancer Society (2005)
6. Kortum, X., Grigull, L., Lechner, W., Klawonn, F.: A dynamic adaptive questionnaire for improved disease diagnostics. In: Adams, N., Tucker, A., Weston, D. (eds.) IDA 2017. LNCS, vol. 10584, pp. 162–172. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68765-0_14
7. Kortum, X., Grigull, L., Muecke, U., Lechner, W., Klawonn, F.: Diagnosis support for orphan diseases: a case study using a classifier fusion method. In: Yin, H., Gao, Y., Li, B., Zhang, D., Yang, M., Li, Y., Klawonn, F., Tallón-Ballesteros, A.J. (eds.) IDEAL 2016. LNCS, vol. 9937, pp. 379–385. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46257-8_41
8. Liaw, A., Wiener, M.: Classification and regression by random forest. R News **2**(3), 18–22 (2002)
9. Ma, L., Liu, X., Song, L., Zhou, C., Zhao, X., Zhao, Y.: A new classifier fusion method based on historical and on-line classification reliability for recognizing common ct imaging signs of lung diseases. Comput. Med. Imaging Graph. **40**, 39–48 (2015)
10. Mücke, U., et al.: Patients experience in pediatric primary immunodeficiency disorders: computerized classification of questionnaires. Front. Immunol. **8**, 384 (2017)
11. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 625–632. ACM (2005)
12. Pohar, M., Blas, M., Turk, S.: Comparison of logistic regression and linear discriminant analysis: a simulation study. Metod. Zv. **1**(1), 143 (2004)
13. Sboner, A., et al.: A multiple classifier system for early melanoma diagnosis. Artif. Intell. Med. **27**(1), 29–44 (2003)
14. Stout, Q.F.: Isotonic regression algorithms. Accessed 6 Aug 2011
15. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 694–699. ACM (2002)