Hypertension Type Classification Using Hierarchical Ensemble of One-Class Classifiers for Imbalanced Data

Bartosz Krawczyk and Michał Woźniak

Department of Systems and Computer Networks,
Wroclaw University of Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{bartosz.krawczyk,michal.wozniak}@pwr.wroc.pl

Abstract. The paper presents the research on the computer support system which is able to recognize the type of hypertension. This diagnostic problem is highly imbalanced, because only ca. 5% of patient suffering from hypertension are diagnosed as secondary hypertension. Additionally the secondary hypertension could be caused by several disorders (in our work we recognize the five most popular reasons) which require strikingly different therapies. Thus, appropriate classification methods, which take into consideration the nature of the decision task should be applied to this problem. We decided to employ the original classification methods developed by our team which have their origin in one-class classification and the ensemble learning. They quality was confirmed in our previous works. The accuracy of the chosen classifiers was evaluated on the basis of the computer experiments which were carried out on the real data set obtained from the hypertension clinic. The results of the experimental investigations confirmed usefulness of the proposed, hierarchical one-class classifier ensemble and could be applied in the real medical decision support systems.

Keywords: classifier ensemble, pattern classification, one-class classifier, imbalanced data, hypertension.

1 Introduction

Medical decision support systems have been focus of intense research for years. According to [1] ca. 11% of expert systems are related to the medical problems, while more than 21% of scientific papers describe applications of machine learning algorithms to medical decision tasks. This work follows-on to our previous works on hypertension type classification [2]. The hypertension is sometimes called *a silent killer*, because many persons do not realize themselves that they suffer from this disorder, which can lead to the serious health problems as coronary heart disease, heart or kidney failure, and stroke to enumerate only a few. Therefore an accurate diagnostic method which could help physician to propose an appropriate therapy is still very desirable tool. Basically, we could

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_34

distinguished two main types of hypertension: the essential hypertension and the secondary one which could be caused by the several disorders. In this work we will use the hypertension taxonomy presented in Fig. 1.

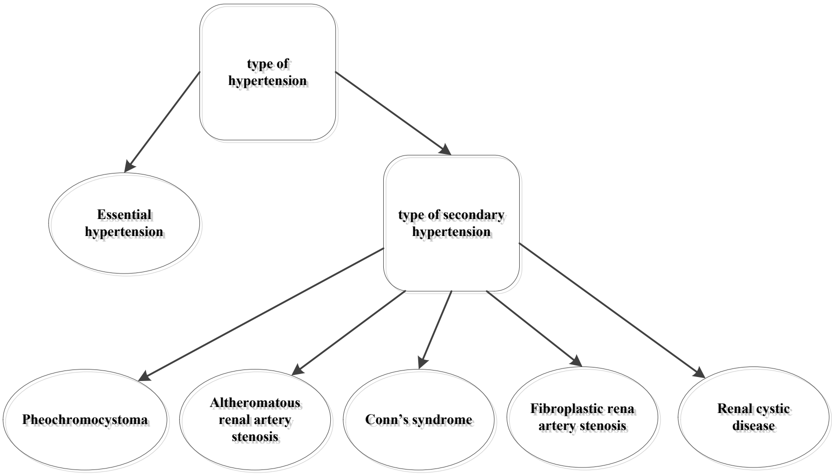

Fig. 1. The taxonomy of the hypertension

According to the medical reports ca. 95% of patient with high blood pressure suffer from the essential hypertension, while rest of them are diagnosed with the one of the five secondary types of hypertension. The problem of hypertension type diagnosis is the crucial stage for the appropriate therapy planning, but as we see it is highly imbalanced what causes that especially patients suffer from the secondary hypertension could be treated incorrectly. Therefore to solve this classification task we should use the classification methods which will take imbalanced data into consideration.

The work is organized as follows. Firstly, the hypertension problem will be described, then classification methods dedicated to imbalance pattern classification task and one-class classification approach are presented shortly. Afterwards, we present the experimental evaluation of chosen classification methods for the problem under consideration. The last section concludes the paper.

2 Hypertension

Blood pressure is determined by the amount of blood the heart pumps and the amount of resistance to blood flow in the arteries. Hypertension is a common condition in which the force of the blood against the artery walls is high enough. The hypertension is sometimes called *a silent killer*, because persons can have it for years without symptoms, but the damage to blood vessels and heart continues. Uncontrolled hypertension strongly increases the risk of health problems, including kidney disorders, coronary heart problems, heart attack, and stroke to enumerate only a few. Normal blood pressure at rest is within the range of 100–140 mmHg systolic and 60–90 mmHg diastolic. Fortunately, the blood pressure measurement is easy and available for the most of the patients, therefore this disorder can be easily detected. Nowadays, the hypertension is recognized as the one of the main health problem, because e.g., [3] reports that in 2013 30-45% of Europeans suffer from it. It has also a huge impact on the global economy, e.g., the American Heart Association estimated the direct and indirect costs of high blood pressure in 2010 as $76.6 billion [4].

The hypertension's therapy planning is usually long and continuous process. The crucial role plays the recognition of its type. Basically, there can distinguish:

– Primary hypertension, known as essential one, that has no known cause and it is diagnosed in the majority of people, i.e., in ca. 95% of hypertension cases.
– Secondary hypertension, which is often caused by comorbid conditions, and is sometimes curable.

The physician is responsible for deciding if the hypertension is of an essential or a secondary type (so called the first level diagnosis). Because only ca. 5% of patients suffering from secondary hypertension, we face with the very hard, highly imbalanced classification problem. Additionally, there are several types of the secondary hypertension. The senior physicians from the *Broussais Hospital of Hypertension Clinic* and *Wroclaw Medical Academy* suggested its following classification:

1. fibroplastic renal artery stenosis,
2. atheromatous renal artery stenosis,
3. Conn's syndrome,
4. renal cystic disease,
5. pheochromocytoma.

3 Imbalanced Classification

Typical classification algorithms work under an assumption that the distribution of objects among the classes in the training set is roughly equal. However, many real-life applications are characterized by the fact, that it is impossible to gather equal number of examples from all classes, as some may appear less frequently or be more costly to gather. In case where one of the classes is represented by a significantly greater number of examples than other, we deal with a problem known as the imbalanced classification. Such an uneven distribution tends to result in a bias of the decision boundary produced by classifiers towards the majority class. This significantly damages the classifier performance on the minority class. With such a situation arises a need for applying carefully designed algorithms that can cope with such a difficult data distribution.

Recent works report that the unequal number of examples among classes is not the major source of problem [5]. In case, where the problem is imbalanced, but the minority class is well-represented by a significant number of objects, even standard algorithms can return good recognition rate [6]. The underlying difficulty is connected to specific data properties, that often are embedded in imbalanced data sets, such as class overlapping [7], small sample size or small disjuncts [8].

These data properties are the major reason behind a poor performance of standard classifiers on imbalanced data sets. Therefore, in recent years there has been a significant development of dedicated methods, that are able to overcome mentioned difficulties. They can be divided into four groups [9]:

- Data-level methods that work at the pre-processing phase of the data. They usually aim at re-balancing the distribution between the classes and they are not dependent on used classifier model. The best known technique here is SMOTE [10], which adds synthetic objects to the minority class.
- Classifier-level methods try to modify the existing classifiers in order to make them robust to unequal object distributions. This is mainly done by reducing or eliminating the bias towards the majority class or shifting the emphasis of the learning step towards the minority class.
- Cost-sensitive approaches introduce a high penalty factor for misclassifying the minority class objects. Instead of standard 0-1 loss function, they use a pre-defined cost matrix, that allows to define the penalty for the learning algorithm for misclassifying minority samples. The most popular are cost-sensitive decision trees, however cost-sensitive neural networks or support vector machines have been also introduced.
- Hybrid approach that uses a classifier ensemble together with one of the mentioned above techniques [11]. They take a full advantage of committee approaches combined with effective method for handling imbalance. Most popular methods include SMOTEBoost [12], EasyEnsemble [13] and AdaCost [14].

4 One-Class Classification

Because in our research we use one-class classification (OCC) approach [15], then let's present this concept shortly. During the training step of OCC only objects from a single class, known as the target concept ω_T, are at disposal. The purpose of OCC is the estimation of a decision surface that encloses all available data samples and thus describes the concept [16]. During the one-class classifier exploitation step, objects from different distributions, unknown during the training phase may appear. They represent data that do not belong to the target concept, and are labeled as outliers ω_O. OCC can be considered as learning in the absence of counterexamples. The target class should be separated from all possible outliers, and hence the decision boundary should be estimated in all directions in the feature space around the target class. This allows us to create a pattern recognition system that is robust to appearance of new classes or lack of representative counterexamples.

OCC is an attractive solution to many real-life problems where data coming from a single class is abundant but other objects are hard or even impossible to obtain such as spam filtering/intrusion detection [17].

Despite the original aim of OCC to work on cases with no access to counterexamples, there is a number of reports that discuss the usefulness of OCC approach in cases, where objects from all of classes are at disposal. This is explained by several attractive properties of OCC, resulting from their different learning procedure. They do not minimize the classification error, but adapt to the features of the target class. They do not use all of the available knowledge from the training set, which results in worse recognition accuracy than multi-class methods for standard problems. However, in case of difficult data sets, OCC can outperform multi-class algorithms, as single-class classifiers are robust to many difficulties embedded in the nature of data. This seems as a very attractive proposal for dealing with imbalanced data sets, and our previous works confirm that using ensembles of OCC can return highly effective recognition systems for uneven class distributions [18, 19].

In this work, we propose to embed the background medical knowledge about the hypertension problem (see Fig. 1) into the process of designing the medical decision support system. The introduced system is realized as a two-step hierarchical architecture, with each step handling a decomposed part of the recognition task.

Step 1. On the upper level of the introduced architecture, we implement a single one-class classifier to distinguish between *essential* and *secondary* hypertension. In our data set, the *essential* hypertension class has significantly larger number of examples than all of the remaining classes, thus leading to an imbalanced problem. We counter this by using a OCC model that is trained on *essential* hypertension class as the target class. All of the remaining five secondary hypertension classes are fused together to create an outlier class. The system outputs a binary value - it can decide that a new object belongs to the target class (*essential* hypertension) or to the outlier class (*secondary* hypertension). In case of the latter decision, the recognition system moves to the second level of architecture, which is able to distinguish between secondary classes. By this, we are able to efficiently handle imbalanced problem at the first step of our architecture.

Step 2. This step aims at distinguishing between one of five types of *secondary* hypertension. This is implemented by decomposing the original multi-class problem with an ensemble of OCC algorithms. Each of the classes is handled by a dedicated one-class classifier, that adjusts to its properties. Then, we use an Error-Correcting Output Codes [20] to reconstruct an original multi-class problem from single-class responses. This allows us to handle difficult multi-class data with efficient decomposition strategy.

5 Experimental Investigations

The aims of the experiment were:

– propose an efficient medical decision support system for automatic diagnosis of hypertension types;
– examine the usefulness of the proposed hierarchical one-class classifier ensemble and compare it to several state-of-the-art methods.

5.1 Set-Up

The initial works on the hypertension type classification system was developed together with *Service d'Informatique Médicale* from the *University Paris VI* [2]. All data was getting from the medical database *ARTEMIS*, which contains the data of the patients with hypertension, whose have been treated in *Hôpital Broussais* in Paris. Although the set of symptoms necessary to correctly assess the existing hypertension is pretty wide, in practice for the diagnosis, results of 18 examinations (which came from general information about patient, blood pressure measurements and basis biochemical data) are used, whose are presented in Tab. 1.

Table 1. Description of features

#	name	#	name
1	sex	10	efffusion
2	body weight	11	artery stenosis
3	high	12	heart failure
4	cigarette smoker	13	palpitation
5	limb ache	14	carotid or lumbar murmur
6	alcohol	15	serum creatinine
7	systolic blood pressure	16	serum potassium
8	diastolic blood pressure	17	serum sodium
9	maximal systolic blood pressure	18	uric acid

The set-up of used classifiers was as follows:

– As a base one-class classifier for the proposed classifier ensemble, we decided to use Support Vector Data Description (SVDD) with RBF kernel and kernel parameters $\sigma = 0.3$, $C = 8$.
– As reference methods, we use:
  - C4.5 decision tree with post-pruning,
  - a multi-class Support Vector Machine (SVM) with RBF kernel and kernel parameters $\sigma = 0.1$, $C = 10$,
  - Random Forest (RandF) ensemble with 120 decision trees .
– The parameter values were established with a grid-search procedure.

– For comparison purposes we use the mentioned classifiers combined with SMOTE preprocessing, in order to counter the imbalance ratio between *essential* class and remaining ones. For SMOTE algorithm, we use 5 neighbors.

All experiments were done with the usage of combined 5x2 cv F test [21] with $\alpha = 0.05$, that allowed to assess the statistical significance of the obtained results.

5.2 Results

The results of the experiments, with the respect to geometric mean (G-mean) values and statistical analysis are given in Tab. 2.

Table 2. Results of experiments on hypertension type classification

Classifier	G-mean	Statistically better than
C4.5	50.92	-
SVM	55.12	C4.5
RandF	57.98	C4.5,SVM
C4.5+SMOTE	67.82	C4.5,SVM,RandF
SVM+SMOTE	68.28	C4.5,SVM,RandF
RandF+SMOTE	70.07	C4.5,SVM,RandF,C4.5+SMOTE,SVM+SMOTE
Hierarchical OCC	75.86	ALL OTHER METHODS

5.3 Discussion of the Results

On the basis of the presented results we may formulate a few interesting observations. The experiments showed, that our hypertension dataset poses a challenge for standard machine learning algorithms, and that it is not a trivial task to achieve a good performance for this problem. However, the output of the experiment proved the quality of the proposed hierarchical ensemble of one-class classifiers. Let us take a closer look into the performance of each methods.

Canonical classifiers deliver highly unsatisfactory results. Neither C4.5, SVM or Random Forest were able to efficiently discriminate between hypertension types. This comes from the fact, that we deal with a multi-class and highly imbalanced problem. If we had used standard accuracy as measure, these classifiers would perform satisfactory. However, G-mean metric allows us to examine their performance with the respect to uneven distribution between classes. And from it we can see, that they fail to properly recognize minority classes (*secondary* hypertension types).

When embedding a dedicated pre-processing method (namely SMOTE algorithm) into these classifiers, we can see a significant rise of the G-mean value. This is because SMOTE inputs artificial instances into minority classes and is able to reduce the classifier's bias towards the majority class. However, one should have in mind a strong limitation of such techniques. With the usage of SMOTE comes the main problem how many of the artificial samples we should generate. Intuition points that best results should be achieved when classes have equal number of objects. Yet with so big disproportion (approximately 9:1) after some repetitions of this algorithm the new objects will be created only on the basis of previously artificially created ones. Therefore it is hard to conclude if so many artificial objects will be representative for the problem.

Our proposed hierarchical one-class ensemble does not suffer from the mentioned limitations. Additionally, it further significantly boosts the recognition rate, which is reflected by the highest G-mean score and backed-up with statistical testing. This shows that OCC can be an efficient tool for handling highly imbalanced data set, despite the fact that it does not use any knowledge about counterexamples. Extending this with a second-level architecture for decomposing multi-class problem with single-class methods allowed us to achieve a very good discrimination between five *secondary* hypertension classes. Combining this approaches into two-level architecture resulted in a robust and effective medical decision support system for hypertension type classification.

6 Conclusions

The paper presents the experimental evaluation of the set of compound classifiers for highly imbalanced multi-class classification task. The considered task was related to crucial problem of hypertension type diagnosis which is recognized as one of the main serious social disease. The proposed method based on one-class classifier ensemble significantly outperforms other considered methods as hybrid approaches used preprocessing as SMOTE. As we mentioned in the previous section, the problem of imbalance data classification is visible not only in disparity among number of training examples represented considered classes, but what maybe more important in the specific data properties. Therefore our future works on decision support systems for the hypertension type classification will look for an appropriate hybrid classifier which is able to take such properties into consideration.

Acknowledgment. This work was supported by The Polish National Science Centre under the grant PRELUDIUM number DEC-2013/09/N/ST6/03504 and by the statutory funds of the Department of Systems and Computer Networks, Wroclaw University of Technology.

References

1. Liebowitz, J.: The Handbook of Applied Expert Systems. Taylor & Francis (1997), `http://books.google.pl/books?id=6DNgIzFNSZsC`
2. Wozniak, M.: Two-stage classifier for diagnosis of hypertension type. In: Maglaveras, N., Chouvarda, I., Koutkias, V., Brause, R. (eds.) ISBMDA 2006. LNCS (LNBI), vol. 4345, pp. 433–440. Springer, Heidelberg (2006)
3. Mancia, G., et al.: esh/esc guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the european society of hypertension (esh) and of the european society of cardiology (esc). European Heart Journal 34(28), 2159–2219 (2013)
4. Lloyd-Jones, D., Adams, R.J., Brown, T.M., Carnethon, M., Dai, S., De Simone, G., Ferguson, T.B., Ford, E., Furie, K., Gillespie, C., et al.: Heart disease and stroke statistics 2010 update a report from the american heart association. Circulation 121(7), e46–e215 (2010)

5. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23(4), 687–719 (2009)
6. Chen, X., Wasikowski, M.: Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 124–132 (2008)
7. Garcia, V., Mollineda, R.A., Sánchez, J.S.: On the k-nn performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications 11(3-4), 269–280 (2008)
8. Napierala, K., Stefanowski, J.: Identification of different types of minority class examples in imbalanced data. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., Cho, S.-B. (eds.) HAIS 2012, Part II. LNCS (LNAI), vol. 7209, pp. 139–150. Springer, Heidelberg (2012)
9. Lopez, V., Fernandez, A., Moreno-Torres, J.G., Herrera, F.: Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. Expert Systems with Applications 39(7), 6585–6608 (2012)
10. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
11. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 42(4), 463–484 (2012)
12. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
13. Liu, X., Wu, J., Zhou, Z.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(2), 539–550 (2009)
14. Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 40(12), 3358–3378 (2007)
15. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54(1), 45–66 (2004)
16. Tax, D., Duin, R.P.W.: Characterizing one-class datasets. In: Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa, pp. 21–26 (2005)
17. Giacinto, G., Perdisci, R., Del Rio, M., Roli, F.: Intrusion detection in computer networks by a modular ensemble of one-class classifiers. Inf. Fusion 9, 69–82 (2008)
18. Krawczyk, B., Woźniak, M.: Diversity measures for one-class classifier ensembles. Neurocomputing 126, 36–44 (2014)
19. Krawczyk, B., Woźniak, M., Cyganek, B.: Clustering-based ensembles for one-class classification. Information Sciences 264, 182–195 (2014)
20. Wilk, T., Wozniak, M.: Soft computing methods applied to combination of one-class classifiers. Neurocomput. 75, 185–193 (2012)
21. Alpaydin, E.: Combined 5 x 2 cv f test for comparing supervised classification learning algorithms. Neural Computation 11(8), 1885–1892 (1999)