machine learning, pattern classification,
class noise, noise filtering,
decision support systems

José A. SÁEZ[1], Bartosz KRAWCZYK[2], Michał WOŹNIAK[2]

HANDLING CLASS LABEL NOISE IN MEDICAL PATTERN CLASSIFICATION SYSTEMS

Pattern classification systems play an important role in medical decision support. They allow to automatize and speed-up the data analysis process, while being able to handle complex and massive amounts of information and discover new knowledge. However, their quality is based on the classification models built, which require a training set. In supervised classification we must supply class labels to each training sample, which is usually done by domain experts or some automatic systems. As both of these approaches cannot be deemed as flawless, there is a chance that the dataset is corrupted by class noise. In such a situation, class labels are wrongly assigned to objects, which may negatively affect the classifier training process and impair the classification performance. In this contribution, we analyze the usefulness of existing tools to deal with class noise, known as noise filtering methods, in the context of medical pattern classification. The experiments carried out on several real-world medical datasets prove the importance of noise filtering as a pre-processing step and its beneficial influence on the obtained classification accuracy.

1. INTRODUCTION

Medical data analysis and medical decision making are acknowledged as important yet difficult tasks and typically based on years of experience of an expert. At the same time, *Computer Aided Diagnosis* (CAD) systems and computer-based data analysis have received increased attention, either to facilitate the work of clinicians or to provide a second opinion [9]. Consequently, there is an increasing interest in well performing pattern classification algorithms for medical data [15].

Machine learning and pattern classification algorithms have gained a significant attention from the medical community in last decade. They allow to quickly process massive, streaming, and complex data, providing a fast and effective automatic diagnosis, even though the final decision always lies in the hands of a physician. However, such an aid is often most valuable, especially for inexperienced physicians or to prevent the presence of routine and the decreased awareness of an expert due fatigue or stress.

In the literature we can find a plethora of reports on successful applications of various machine learning methods to medical domains. In the last years, there is an intense focus on contemporary trends in classification systems transposed to clinical decision support. One

[1]ENGINE Centre, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland, e-mail: jose.saezmunoz@pwr.edu.pl

[2]Department of Systems and Computer Networks, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland, e-mail:{bartosz.krawczyk,michal.wozniak}@pwr.edu.pl

should pay special attention to recent developments in applications of fuzzy classifiers [22], extreme learning [4], ensemble systems [11], data streams [7], or big data analytics [1].

When designing a medical decision support system we must take into consideration the following factors:

- **Learning set**. This is the initial step of the system design. One needs to collect a set of examined examples that will serve as a basis for the learning process. Such a set must consist of a representative sample of objects that are described by properly extracted features of high discriminative power. Missing values should be avoided, samples from all of the possible classes should be included, and examples should well describe the possible instances of the examined medical case. Finally, one needs a domain expert or an automatic procedure to provide a class label to each case that will be utilized by the classification method.
- **Classification method**. When obtaining a proper learning set, one must select a pattern classifier that will be able to efficiently extract the properties of data. The data will be used to construct decision rules or a separating hyperplane, which are employed in the decision making process. The classifier must be able to generalize the knowledge embedded into the training set onto new, unseen examples. The classifier selection step must be performed to choose the most efficient learner for each dataset, as according to *No Free Lunch Theorem* [26] there is no single universal classification method.
- **Ease of use**. The created medical decision support system will be used not by machine learning experts, but by physicians. Therefore, it should be either delivered as a black box solution or be free of so-called *magic parameters* that must be tuned to the specific problem. Additionally, many physicians prefer the decision to be interpretable, as it may shed additional light on the patient's condition.

Each of these components is equally important. In this contribution, we will focus on the first one, specifically on the quality of the provided class labels within the dataset. We assume that the dataset is already collected and might be plagued by the so-called class (or label) noise [6]. In such a scenario some of the objects were mislabeled by the expert and thus they will mislead the classifier training procedure.

Among the existing tools to deal with class noise, noise filters, which remove noisy examples from the training data, are widely used due to their benefits in the learning in terms of classification accuracy [3], [19]. We propose to examine several of such noise filtering methods regarding to their usefulness for handling varied levels of class noise. We apply them on a set of real-world medical datasets to show the practical importance of the proposed pre-processing analysis. We verify these algorithms on the basis of computer experiments that allow us to draw conclusions regarding to the danger of class noise in clinical decision support systems.

2. NOISY DATA IN MEDICAL PATTERN CLASSIFICATION

The presence of noise in data is a common problem that produces several negative consequences in classification problems. In multi-class problems, these consequences are aggravated in terms of accuracy, building time, and complexity of the classifiers. The classification accuracy of a learner is directly influenced by the quality of the training data used. Real-world datasets usually contain corrupted data that may hinder the interpretations, decisions, and therefore, the classifiers built from that data.

Class noise occurs when an example is incorrectly labeled [17]. Class noise can be attributed to several causes, such as subjectivity during the labeling process, data entry errors, or inadequacy of the information used to label each example. Two types of class noise can be distinguished [18]:

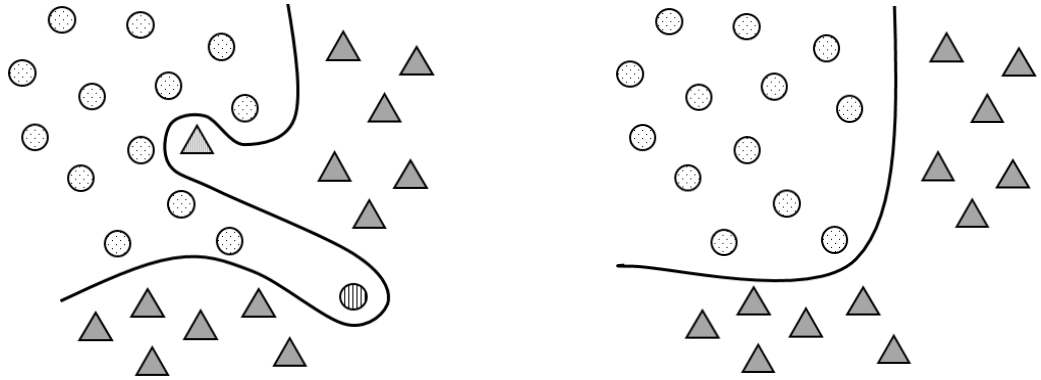

Fig. 1. Toy dataset with two examples contaminated by class noise. The first figure shows the exemplary decision boundary created when considering the noisy samples as proper ones. The second figure shows a decision boundary after the removal of the noisy examples during the pre-processing filtering. One can see that the decision boundary obtained after the filtering is less complex and will lead to a better generalization over these two classes.

- contradictory examples - duplicate examples described by different class labels.
- mislabeled examples - examples assigned to another class than the real one.

An example of the behavior of classifiers in presence of noisy samples is given in Figure 2.

In medical pattern classification, class noise may originate from several sources, among which the most possible are:

- *Human errors*. These may happen quite often, especially when an expert physician is asked to provide labels for a large number of examples. Mistakes may occur due to weariness, routine, quick examination of each case, time pressure, or even due to not paying attention to potential outliers or atypical cases. Additionally, especially in case of highly complex data, we cannot assume that the physician will be infallible.
- *Machine errors*. These is especially present in cases, where a machine is responsible for providing automatic labels. Here some design faults, momentary error or too similar cases may lead to the presence of erroneous labels.
- *Digitalization errors*. When creating a digital record of the examined cases, one may simply incorrectly input a class by a mistake.
- *Archiving errors*. When using historical recordings, there is a chance of missing or incorrectly copied information.

Wrongly labeled medical cases will damper the effectiveness of pattern classifiers that used them for training. In medical decision support, this fact can affect the final accuracy, which is of high risk as we are often dealing with human health and life. Therefore, filtering the class noise must be of high importance and should not be omitted.

3. EXAMINED NOISE FILTERING METHODS

The noise filters used have been chosen due they apply different filtering procedures and are well-known representatives of the field. The parameter setup for all the noise filters is the default one recommended by the authors of such filters. They are briefly described in the following:

1) *Ensemble Filter* (EF) [3]. EF classifies the training data using an n-fold cross-validation with several classification algorithms (C4.5 [16], 1-NN [14] and LDA [13]). Then, the noisy examples are identified using a voting scheme and removed from the training data.
2) *Edited Nearest Neighbor* (ENN) [24]. This algorithm removes those examples which class does not agree with that of the majority of its k ($k = 3$ in our experiments) nearest neighbors.

3) *Iterative-Partitioning Filter* (IPF) [8]. IPF removes noisy examples in multiple iterations. In each iteration, the training data is split into n folds and C4.5 is built over each of these subsets to evaluate all the examples. Then, the examples misclassified by the majority of the classifiers are removed and a new iteration is started.
4) *Multiedit* (ME) [5]. This splits the training data into n folds. 1-NN classifies the examples from the part x considering the part *(x+1) mod n* as training set and the misclassified examples are removed. This process is repeated until no examples are eliminated.
5) *Nearest Centroid Neighbor Edition* (NCNE) [20]. This is a modification of ENN, which consists of discarding from the training set every example misclassified by the k ($k = 3$) nearest centroid neighbors (k-NCN) rule.
6) *Relative Neighborhood Graph* (RNG) [21]. This technique builds a proximity undirected graph $G = (V, E)$, in which each vertex corresponds to an instance from the training set. There is a set of edges E, so that (x_i, x_j) belongs to E if and only if x_i and x_j satisfy a *neighborhood relation*. In this case, we say that these instances are graph neighbors. The graph neighbors of a given point constitute its graph neighborhood. The filtering scheme discards those instances misclassified by their graph neighbors.

4. EXPERIMENTAL STUDY

We propose a set of experiments carried-out on real-world medical datasets. We introduce different noise levels into each of them and investigate the accuracy obtained by a given classifier with the respect to the increase of the class noise ratio. We apply the filters described in Section 3 to see how significantly they can alleviate the negative impact of wrongly labeled training examples on the final accuracy.

4.1. DATASETS

We selected four real medical datasets describing different types of applications of clinical decision support systems. Their details are given in Table 1.

1) *Cytological slide examination* (breastcancer) [10]. This dataset describes the automatic examination of breast cancer cytology slides taken with fine needle biopsy. Firefly algorithm is applied to raw images in order to detect the position of nuclei. This is used to initialize the watershed algorithm that allows for an efficient segmentation of cells. A set of features describing morphological properties of these nuclei is then extracted. Three types of cancerous cells are distinguished.
2) *Hypertension detection* (hypertension) [12]. This is a multi-class imbalanced dataset that deals between the diagnosis of first order hypertension and five types of secondary hypertensions. Features are extracted from a set of medical records and examinations.
3) *Chronic kidney disease* (kidneydisease)[1]. This deals with the early detection of kidney failure and it was collected from nearly 2 months of period in an hospital.
4) *Mice protein expression* (miceprotein)[2]. This dataset contains expression levels of 77 proteins measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning.

[1]https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
[2]https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression

Table 1. Details of the datasets.

Dataset	Objects	Features	Classes	Objects per class
breastcancer	675	36	3	225, 225, 225
hypertension	1425	18	6	77, 147, 912, 113, 87, 89
kidneydisease	400	25	2	250, 150
miceprotein	1080	82	8	150, 150, 135, 135, 135, 135, 105, 135

4.2. SET-UP

Some of the datasets (*hypertension* and *kidneydisease*) contains missing values. In order to avoid their influence in classifier performance and focus on class noise's effects we must treat these values. The two most common approaches in this scenario are either removing examples containing missing values or filling them with imputation methods [2]. In order to not reduce still more the number of examples in the final datasets (since noise filters also remove examples from the training data), we decide to use the second approach. For this reason, we first employ the k-*Nearest Neighbor Imputation* (KNNI) [2] method to fill the missing values. Based on the k-NN algorithm, every time a missing value is found in a current example, KNNI computes the k nearest neighbors ($k = 10$ in our experiments) and their average value is imputed. KNNI uses the HVDM distance [25], which is valid for both nominal and numerical attributes.

Then, in order to control the amount of noise in each of the four datasets, different noise levels $x\%$ are introduced into each training dataset in a supervised manner following an *uniform class noise scheme* [23]: $x\%$ of the examples are corrupted randomly replacing the class labels of these examples by other ones from the set of classes. Thus, using this scheme any of the classes of the dataset may be affected by the noise. We will consider the noise levels $x = 0\%$ (base datasets, without additional noise), $x = 5\%$, $x = 10\%$, $x = 20\%$ and $x = 30\%$.

The accuracy estimation of a classifier in a dataset is obtained by means of a 5-fold stratified cross-validation. In order to create a noisy dataset from the original one, the noise is introduced into the training partitions as follows:

1) A level of noise $x\%$ of class noise is introduced into a copy of the full original dataset.
2) Both datasets, the original one and the noisy copy, are partitioned into 5 equivalent folds, that is, with the same examples in each one.
3) The training partitions are built from the noisy copy, whereas the test partitions are formed from examples from the base dataset.

Finally, we use the C4.5 [16] decision tree as a base classifier. We report the results according to the obtained classification accuracy.

4.3. RESULTS AND DISCUSSION

The experimental results with respect to different noise levels are given in Table 2.

The experimental results obtained allow us to draw some interesting conclusions regarding the stated problem of noisy medical data classification.

Let us firstly concentrate on the performance of C4.5 without and with the presence of the noise. For datasets without additional noise, we obtain a satisfactory accuracy that shows the potential usefulness of machine learning for the stated problem. However, with the occurrence of noise and increase of its ratio, we can see a quick deterioration of the obtained classification accuracy. Even as small as 5% of noisy samples already affects the classifier (as in *miceprotein*, where we observe more than 3.5% of accuracy drop). For higher noise ratios, C4.5 without any pre-processing becomes highly inaccurate (as in *breastcancer* dataset, where we can observe almost 20% of drop in accuracy). Therefore, using noise filtering as a pre-processing step becomes a mandatory task in designing such a medical pattern classification system.

Table 2. Accuracies [%] obtained by C4.5 with respect to a given class noise level and noise filtering method.

Dataset	Noise	None	EF	ENN	IPF	ME	NCNE	RNG
breastcancer	0	90.37	89.48	88.30	90.37	76.74	**91.26**	90.67
	5	90.67	90.22	87.26	90.81	76.15	**91.85**	90.52
	10	84.15	88.44	86.07	87.56	76.74	**90.37**	86.67
	20	76.74	86.07	86.96	**87.70**	74.52	86.37	83.70
	30	71.85	**87.70**	79.41	86.52	73.78	87.56	79.56
hypertension	0	65.47	**69.47**	66.25	68.91	64.63	68.49	65.33
	5	65.47	**69.89**	67.58	68.35	64.63	68.35	66.18
	10	63.93	**69.12**	66.18	68.07	64.84	67.44	65.40
	20	63.09	**68.56**	65.47	67.72	64.56	67.16	66.81
	30	54.67	**68.63**	65.19	67.79	64.91	66.04	63.44
kidneydisease	0	96.50	95.75	94.25	95.25	95.25	**97.25**	94.50
	5	**96.50**	95.00	**96.50**	94.75	92.00	**96.50**	94.50
	10	93.50	95.50	91.25	95.50	95.00	**96.75**	93.50
	20	95.00	94.00	95.25	94.50	87.00	**95.50**	92.75
	30	91.00	94.25	88.25	**95.00**	91.50	94.25	89.50
miceprotein	0	**85.28**	83.52	81.30	78.06	76.76	85.19	83.15
	5	81.57	**83.70**	82.13	76.39	76.30	82.41	81.39
	10	77.87	**81.94**	81.39	77.04	72.59	81.39	79.81
	20	72.87	78.52	**81.57**	76.57	70.83	79.91	77.69
	30	64.44	73.33	74.81	72.04	64.91	**75.09**	73.70

Using filtering can significantly alleviate the negative role of mislabeled samples (as in the aforementioned *breastcancer* dataset, where filtering leads to over 15% of accuracy gain).

What is highly interesting, for some of the datasets without additional noise, using some noise filters still leads to small improvements in accuracy. This can be explained by the fact that these datasets are already affected by some small degree of class noise on their own. This puts further emphasis on the usage of noise filtering in the process of medical pattern classification.

There is also one case without additional noise, for the *miceprotein* dataset, in which none of the noise filters improve the performance of C4.5 without preprocessing. This fact can be attributed to higher amounts of class noise in the data allow the noise filters to better show their potential. Furthermore, one must note that noise filters do not always provide an improvement in the performance results [19] and this depends on many components, such as the characteristics of the data treated.

When carrying out a comparison among the examined filtering methods we may easily observe that there are two dominant approaches: EF and NCNE. They both display excellent robustness to noise and are able to efficiently handle even highly contaminated datasets (30% of noisy objects in the training set). One should note that the differences between examined filters are in most cases relatively small and all of them allow to alleviate the influence of noisy samples. However, both EF and NCNE filters offer the highest flexibility and robustness, and thus are the two recommended methods to apply in medical pattern classification systems.

5. CONCLUSIONS

In this contribution we have discussed the role of the quality of class labels provided in the training set for designing a medical decision support system based on machine learning. Class noise may be a result of human expert error when describing the given cases or a result from some erroneous data gathering. Incorrectly labeled samples may strongly influence the learning process, leading to creation of complex decision boundaries that have reduced generalization abilities. Therefore, the proper handling of such uncertain objects is of high importance when constructing pattern classifiers.

We have examined the usefulness of six different noise filters in the presence of varied

ratios of noise. The experimental study carried out on a set of real-world medical datasets prove that even small levels of class noise may significantly decrease the classifiers' quality. Noise filtering is especially crucial in cases in which a significant amount of training objects may suffer from wrong labeling.

In future we plan to extend our works on noise filtering to massive and streaming medical datasets.

ACKNOWLEDGEMENT

José A. Sáez was supported by EC under FP7, Coordination and Support Action, Grant Agreement Number 316097, ENGINE European Research Centre of Network Intelligence for Innovation Enhancement (http://engine.pwr.wroc.pl).

Bartosz Krawczyk and Michał Woźniak were supported by the Polish National Science Center under the grant no. DEC-2013/09/B/ST6/02264.

BIBLIOGRAPHY

[1] AZAR A. T., HASSANIEN A. E. Dimensionality reduction of medical big data using neural-fuzzy classifier. Soft Comput., 2015, Vol. 19. pp. 1115–1127.

[2] BATISTA G. E. A. P. A., MONARD M. C. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 2003, Vol. 17. pp. 519–533.

[3] BRODLEY C. E., FRIEDL M. A. Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research, 1999, Vol. 11. pp. 131–167.

[4] CZARNECKI W. M. Weighted tanimoto extreme learning machine with case study in drug discovery. IEEE Comp. Int. Mag., 2015, Vol. 10. pp. 19–29.

[5] DEVIJVER P. On the editing rate of the MULTIEDIT algorithm. Pattern Recognition Letters, 1986, Vol. 4. pp. 9–12.

[6] GARCIA L. P. F., DE CARVALHO A. C. P. L. F., LORENA A. C. Effect of label noise in the complexity of classification problems. Neurocomputing, 2015, Vol. 160. pp. 108–119.

[7] HUANG G., ZHANG Y., CAO J., STEYN M., TARAPOREWALLA K. Online mining abnormal period patterns from multiple medical sensor data streams. World Wide Web, 2014, Vol. 17. pp. 569–587.

[8] KHOSHGOFTAAR T. M., REBOURS P. Improving software quality prediction by noise filtering techniques. Journal of Computer Science and Technology, 2007, Vol. 22. pp. 387–396.

[9] KONONENKO I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 2001, Vol. 23. pp. 89–109.

[10] KRAWCZYK B., FILIPCZUK P. Cytological image analysis with firefly nuclei detection and hybrid one-class classification decomposition. Engineering Applications of Artificial Intelligence, 2014, Vol. 31. pp. 126–135.

[11] KRAWCZYK B., SCHAEFER G. A hybrid classifier committee for analysing asymmetry features in breast thermograms. Appl. Soft Comput., 2014, Vol. 20. pp. 112–118.

[12] KRAWCZYK B., WOŹNIAK M. Hypertension type classification using hierarchical ensemble of one-class classifiers for imbalanced data. ICT Innovations 2014, 2015, Vol. 311 of Advances in Intelligent Systems and Computing. pp. 341–349.

[13] LE CESSIE S., VAN HOUWELINGEN J. Ridge estimators in logistic regression. Applied Statistics, 1992, Vol. 41. pp. 191–201.

[14] MCLACHLAN G. J. Discriminant Analysis and Statistical Pattern Recognition (Wiley Series in Probability and Statistics). 2004. Wiley-Interscience.

[15] POMBO N., ARAÚJO P., VIANA J. Knowledge discovery in clinical decision support systems for pain management: A systematic review. Artificial Intelligence in Medicine, 2014, Vol. 60. pp. 1–11.

[16] QUINLAN J. R. C4.5: programs for machine learning. 1993. Morgan Kaufmann Publishers, San Francisco, CA, USA.

[17] SÁEZ J. A., GALAR M., LUENGO J., HERRERA F. Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition. Knowl. Inf. Syst., 2014, Vol. 38. pp. 179–206.

[18] SÁEZ J. A., GALAR M., LUENGO J., HERRERA F. INFFC: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control. Information Fusion, 2016, Vol. 27. pp. 19–32.

[19] SÁEZ J. A., LUENGO J., HERRERA F. Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification. Pattern Recognition, 2013, Vol. 46. pp. 355–364.

[20] SÁNCHEZ J., BARANDELA R., MÁRQUES A., ALEJO R., BADENAS J. Analysis of new techniques to obtain quality training sets. Pattern Recognition Letters, 2003, Vol. 24. pp. 1015–1022.

[21] SÁNCHEZ J., PLA F., FERRI F. Prototype selection for the nearest neighbor rule through proximity graphs. Pattern Recognition Letters, 1997, Vol. 18. pp. 507–513.

[22] SANZ J., GALAR M., JURIO A., BRUGOS A., PAGOLA M., BUSTINCE H. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. Appl. Soft Comput., 2014, Vol. 20. pp. 103–111.

[23] TENG C.-M. Correcting Noisy Data. Proceedings of the Sixteenth International Conference on Machine Learning, 1999. Morgan Kaufmann Publishers, San Francisco, CA, USA, pp. 239–248.
[24] WILSON D. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems and Man and Cybernetics, 1972, Vol. 2. pp. 408–421.
[25] WILSON D. R., MARTINEZ T. R. Improved heterogeneous distance functions. Journal of Artificial Intelligence Research, 1997, Vol. 6. pp. 1–34.
[26] WOLPERT D. The supervised learning no-free-lunch theorems. In Proc. 6th Online World Conference on Soft Computing in Industrial Applications, 2001. pp. 25–42.