Combining Cluster Validation Indices for Detecting Label Noise

Veselka Boeva, Jan Kohstall, Lars Lundberg and Milena Angelova

Abstract In this paper, we show that cluster validation indices can be used for filtering mislabeled instances or class outliers prior to training in supervised learning problems. We propose a technique, entitled Cluster Validation Index (CVI)-based Outlier Filtering, in which mislabeled instances are identified and eliminated from the training set, and a classification hypothesis is then built from the set of remaining instances. The proposed approach assigns each instance several cluster validation scores representing its potential of being an outlier with respect to the clustering properties the used validation measures assess. We examine CVI-based Outlier Filtering and compare it against the Local Outlier Factor (LOF) detection method on ten data sets from the UCI data repository using five well-known learning algorithms and three different cluster validation indices. In addition, we study and compare three different approaches

Veselka Boeva · Jan Kohstall[†] · Lars Lundberg
Blekinge Institute of Technology, SE-371 79, Karlskrona, Sweden
🖂 veselka.boeva@bth.se
🖂 lars.lundberg@bth.se
🖂 jan.kohstall@bth.se

[†] Jan Kohstall is also affiliated with Hasso Plattner Institute, University of Potsdam (Germany).

Milena Angelova
Technical University of Sofia, Plovdiv, Bulgaria
🖂 mangelova@tu-plovdiv.bg

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/18
ISSN 2363-9881

for combining the selected cluster validation measures. Our results show that for most learning algorithms and data sets, the proposed CVI-based outlier filtering algorithm outperforms the baseline method (LOF). The greatest increase in classification accuracy has been achieved by using union or ranked-based median strategies to assemble the used cluster validation indices and global filtering of mislabeled instances.

1 Introduction

Supervised learning algorithms are used to generate classifiers (Kohavi, 1995). For this machine learning task, the main idea is to apply a learning algorithm to detect patterns in a data set (inputs) that are associated with known class labels (outputs) in order to automatically create a generalization; i.e., a classifier. However, noise and outliers exist in real world data sets due to different errors. When the data is modeled using machine learning algorithms, the presence of label noise and outliers can affect the generated model. Improving how learning algorithms handle noise and outliers can produce better models.

Outlier mining is the process of finding unexpected events and exceptions in the data. There is a lot of work on outlier detection including statistical methods (Kubica and Moore, 2003), rule creation (Khoshgoftaar et al, 2004). Conventional outlier mining methods find exceptions or rare cases with respect to the whole data set.

In this paper, we introduce a novel outlier filtering technique that is close to class outlier detection approaches which find suspicious instances taking into account the class label (He et al, 2004; Hewahi and Saad, 2007; Papadimitriou and Faloutsos, 2003). Such filtering approaches are also referred to as label noise cleansing (Frénay and Verleysen, 2014). The proposed approach, called Cluster Validation Index (CVI)-based Outlier Filtering, applies cluster validation measures to identify mislabeled instances or class outliers. We remove these instances prior to training and study how this affects the performance of the machine learning algorithm.

Cluster validation measures are usually used for evaluating and interpreting clustering solutions in unsupervised learning. In our approach we use them for detecting outliers in supervised learning scenarios. The intuition behind our strategy is that instances in the training set that are not strongly connected to their clusters are mislabeled instances or class outliers and should be removed prior to training to improve the generated classifier.

We assign each instance in the training set several cluster validation scores representing its potential of being a class outlier with respect to the clustering properties the used validation measures assess. In this respect, the proposed approach may be referred to as a multi-criteria outlier filtering measure. Namely, it uses a combination of different cluster validation indices (Silhouette Index (SI), Connectivity (Co) and Average Intracluster gap (IC-av)) in order to reflect different aspects of the clustering model determined by the labeled instances of the training set.

The CVI-based outlier filtering is compared against the Local Outlier Factor (LOF) detection method (Breunig et al, 2000), a well-known baseline and outlier detection algorithm. Our results reveal that the proposed CVI-based outlier filtering approach outperforms the used baseline algorithm (LOF) for most of the experimental scenarios.

The work presented in this paper is an extended study based on the initial results published in Boeva et al (2018). In the current work, we have studied and validated an additional cluster validation indices' combination technique.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 discusses the cluster validation measures and describes the proposed class outlier filtering approach. Section 4 presents the evaluation of the discussed approach. Section 5 is devoted to conclusions and future work.

2 Related Work

A number of methods that treat individual instances in a data set differently during training to focus on the most informative ones have been developed. For example, an automated method that orders the instances in a data set by complexity based on their likelihood of being misclassified for supervised classification problems is presented in Smith and Martinez (2016). The underlying assumption of this method is that instances with a high likelihood of being misclassified represent more complex concepts in a data set. The authors Brodley and Friedl (1999), Gamberger et al (2000), and Smith and Martinez (2011) have shown that focusing on the simpler instances during training significantly increases generalization accuracy. Identifying and removing noisy instances and outliers from a data set prior to training generally results in an increase in classification accuracy on non-filtered test data.

Conventional outlier detection methods find exceptions or rare cases in a data set irrespective of the class label of these cases, whereas class outlier detection approaches find suspicious instances taking the class label into account as described by He et al (2002), He et al (2004), Hewahi and Saad (2007), and Papadimitriou and Faloutsos (2003).

Closely related to class outlier mining is noise reduction by Segata and Blanzieri (2010) and Tomek (1976) that attempts to identify and remove mislabeled instances. For example, Brodley and Friedl (1999) attempt to identify mislabeled instances using an ensemble of classifiers. Rather than determining if an instance is mislabeled, the approach introduced in Smith and Martinez (2011) filters instances that should be misclassified. A comprehensive survey on the different types of label noise, their consequences and the algorithms that consider label noise is presented by Frénay and Verleysen (2014). In addition, a review of some typical problems associated with high-dimensional data and outlier detection specialized for high-dimensional data is published in Zimek et al (2012).

To improve the detection rate of outlier methods several other authors proposed outlier ensembles, e.g., Aggarwal (2013) gives a good overview. Furthermore several ranked based methods for combining cluster validation measures are proposed by Vendramin et al (2013).

3 Methods and Technical Solutions

3.1 Cluster Validation Techniques

One of the most important issues in cluster analysis is the validation of clustering results. Essentially, the cluster validation techniques are designed to find the partitioning that best fits the underlying data, and should therefore be regarded as a key tool in the interpretation of clustering results. The data mining literature provides a range of different cluster validation measures, which are broadly divided into two major categories: *External* and *internal* as described by Jain and Dubes (1988).

External validation measures have the benefit of providing an independent assessment of clustering quality, since they evaluate the clustering result with respect to a pre-specified structure. However, previous knowledge about data is rarely available. Internal validation techniques, on the other hand, avoid the need for using such additional knowledge, but have the problem that they need to base their validation on the same information used to derive the clusters themselves.

Internal measures can be split with respect to the specific clustering property they reflect and assess to find an optimal clustering scheme, e.g., *compactness*, *separation*, *connectedness*, and *stability* of the cluster partitions. *Compactness* evaluates the cluster homogeneity that is related to the closeness within a given cluster. *Separation* demonstrates the opposite trend by assessing the degree of separation between individual groups. The third type of internal validation measure, *connectedness*, quantifies to what extent the nearest neighboring data items are placed into the same cluster. The *stability* measures evaluate the consistency of a given clustering partition by clustering from all but one experimental condition. The remaining condition is subsequently used to assess the predictive power of the resulting clusters by measuring the within-cluster similarity in the removed experiment.

Thus in this work, we have decided to use three internal validation measures for analyzing the labeled instances prior to training in supervised classification problems in order to identify mislabeled ones. Based on the above mentioned classification, we have selected one validation measure for assessing compactness and separation properties of a partitioning (*Silhouette Index*, SI), one for assessing connectedness (*Connectivity*, Co) and one for assessing tightness and dealing with arbitrary shaped clusters (*IC-av*).

The *Silhouette Index* (Rousseeuw (1987)) is a cluster validity index that is used to judge the quality of any clustering solution $C = C_1, C_2, \ldots, C_k$. Suppose a_i represents the average distance of object i from the other objects of the cluster to which the object is assigned, and b_i represents the minimum of the average distances of object i from objects of the other clusters. Then the *Silhouette Index* of object i can be calculated by

$$s(i) = (b_i - a_i)/\max\{a_i, b_i\}. \qquad (1)$$

The overall Silhouette Index for clustering solution C of m objects, is defined as:

$$s(C) = \frac{1}{m}\sum_{i=1}^{m}(b_i - a_i)/\max\{a_i, b_i\}. \tag{2}$$

The values of *Silhouette index* vary from -1 to 1 and a higher value indicates better clustering results.

Connectivity captures the degree to which objects are connected within a cluster by keeping track of whether the neighboring objects are put into the same cluster as introduced by Handl et al (2005). Define m_{ij} as the jth nearest neighbor of object i, and let $\chi_{im_{ij}}$ be zero if i and m_{ij} are in the same cluster and $1/j$ otherwise. Then for a particular clustering solution $C = C_1, \ldots, C_k$ of m objects and a neighborhood size n_r, the *Connectivity* is defined as

$$Co(C) = \sum_{i=1}^{m}\sum_{j=1}^{n_r}\chi_{im_{ij}}. \tag{3}$$

The *Connectivity* has a value between zero and $\sum_1^{n_r} 1/n_r$ and should be minimized. Evidently, the *Connectivity* of object i can be calculated by

$$Co(i) = \sum_{j=1}^{n_r}\chi_{im_{ij}}. \tag{4}$$

IC-av, developed by Bayá and Granitto (2013), estimates cluster *tightness*, but instead of assuming spherical shape, it assumes that clusters are connected structures with arbitrary shape. The connections between the nodes are obtained by using a minimum spanning tree. Then for a particular clustering solution $C = C_1, \ldots, C_k$, the *IC-av* is defined as

$$IC\text{-}av(C) = \sum_{r=1}^{k}\frac{1}{n_r}\sum_{i,j\in C_r} d_{ij}^2, \tag{5}$$

where n_r is the number of objects in cluster C_r $(r = 1, 2, \ldots, k)$ and d_{ij} is the maximum edge distance which represents the longest edge in the path joining objects i and j in the minimum spanning tree (MST) built on the clustered set of objects. The *IC-av Index* of object i, which is partitioned in cluster C_r, can be calculated by

$$IC\text{-}av(i) = \frac{1}{n_r} \sum_{j \in C_r} d_{ij}^2. \tag{6}$$

The *IC-av* has a value between zero and the longest edge in the MST and should be minimized.

3.2 Combining Cluster Validation Indices for Detecting Label Noise

In this study, we propose a class outlier filtering technique, named Cluster Validation Index (CVI)-based Outlier Filtering, that combines cluster validation measures to identify mislabeled instances.

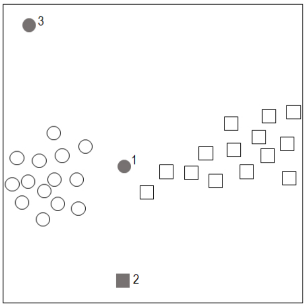

Figure 1: A hypothetical 2-dimensional data set.

We use three internal validation measures for the evaluation of the labeled instances prior to training: Silhouette Index (SI), Connectivity and IC-av. Figure 1 shows a hypothetical 2-dimensional data set with two classes (circle and square) and three outliers (filled in two circles and one square). If we apply SI for assessing the instances of this data set instance 2 will be recognized as an outlier, while instance 1 will be removed in case of Connectivity is used. However, outlier instance 3 will be not considered as an outlier with respect to the SI and Connectivity measures. This instance would be filtered out as an outlier by IC-av measure estimating cluster tightness. The choice of cluster validation measure is therefore crucial for the performance of the proposed outlier mining technique.

According to Bezdek and Pal (1998), a possible approach to bypass the selection of a single cluster validity criterion is to rely on multiple criteria in order to obtain more robust evaluations. In a recent work, Jaskowiak et al (2016) also proposed a method for combining internal cluster validation measures into ensembles, which show superior performance when compared to any single ensemble member. Consequently, a rather straightforward solution to the above described problem is to use different cluster validation measures in order to find some complementarity among the clustering properties they assess. In this way different aspects of the clustering model determined by the known class labels will be reflected in the filtering phase. For instance, the selected cluster validation measures can be combined by logical operators: $\vee$ (OR) or $\wedge$ (AND). We have initially validated this idea in Boeva et al (2018). In the current study, we further evaluate the proposed CVI-based Outlier Filtering algorithm and also study a rank-based median strategy, proposed by Jaskowiak et al (2016), for combining the selected single cluster validation measures in a guided way. The obtained results are further benchmarked to ones generated by applying the above mentioned logical operators for combining the selected CVIs.

4 Evaluation and Results

4.1 Experimental Setup

We study how filtering of mislabeled instances affects the classification accuracy of 10 data sets from the UCI data repository and 5 learning algorithms trained with and without filtering. The algorithms that have been used are: 1 nearest neighbor (1-NN), 5 nearest neighbor (5-NN), Support Vector Machine (SVM), Gaussian Naïve Bayes (GNB), Decision Tree (CART). No parameter optimization has been performed on any of the algorithms.

In this work, we specially study and compare four label noise filtering methods: SI $\wedge$ Co $\wedge$ IC-av, SI $\vee$ Co $\vee$ IC-av, Median(SI, Co, IC-av) and LOF. The first two methods use logical operators to combine the three CVIs, the third method is based on the rank-based median strategy of Jaskowiak et al (2016). The performance of these three CVIs-based class outlier filtering approaches is further evaluated by comparing it against the LOF method of Breunig et al (2000), which is a widely used baseline algorithm.

Each label noise filtering method has been evaluated using 5 by 10-fold cross-validation (running 10-fold cross-validation 5 times, each time with a different seed to partition the data). In each iteration we obtain a training set and a test set. The filtering is performed only on the training set. After that each learning algorithm is trained on the filtered training set. We use the unfiltered test set to evaluate the models.

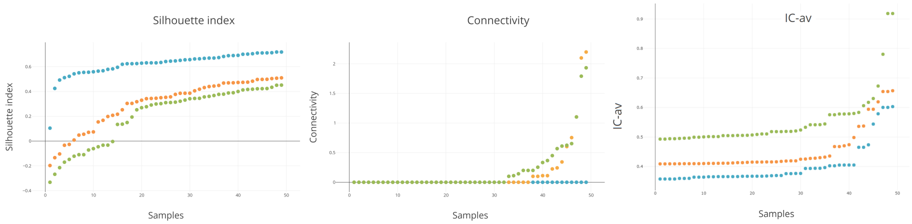

Figure 2: SI, Connectivity and IC-av scores generated on the instances of Iris data set (🟦 Class 0 *Setosa*, 🟧 Class 1 *Versicolor*, 🟩 Class 2 *Virginica*).

Initially, SI, Connectivity and IC-av scores are calculated for each instance of the considered data set. In case of using the logical operators to combine the three CVIs the instances in each class are then ranked based on the assigned cluster validation scores separately for each measure. This is illustrated in Figure 2, which depicts the ranked SI, Connectivity and IC-av scores calculated on the instances of *Iris* data set. In case, the rank-based median strategy is applied to combine the CVIs then the instances in each class are ranked based on the generated median rank. Notice that in the considered context a local (w.r.t. the classes) or global (w.r.t. the entire data set) percentage of the top ranked instances can be identified and filtered out from the training set as outliers. In case of local filtering we remove *x* percent from each class; in global filtering it is enough that we filter *x* percent from the entire training set. In the first case one and the same number of instances with the lowest SI (respectively, the highest Connectivity or IC-av) scores will be removed from each class of the training set. However, the fact that the number of instances identified to be removed as outliers is fixed to be the same for each class may lead to a somewhat random choice of outliers. For instance, it may happen that the SI scores of some instances recognized as outliers are rather high, since these have only been included in the list of outliers in order to reach the required fixed number of instances. This can easily be noticed in Figure 2, e.g., see Class 0 Setosa (Silhouette Index). The described negative effect due to the use of local filtering can be mitigated by applying global filtering, instead. Namely, in this case a varying number of instances is removed as outliers from each class, since these are identified by a percent from the entire training set.

For each experimental scenario we have tested the following (local and global) percentages: 0 %, 2 %, 4 %, 6 %, 8 %, 10 %. For our experiments we have used the following datasets from the UCI repository: digits, ecoli, iris, wine, yeast, breast cancer, arrhythmia, dermatology, white wine quality and red wine quality.

As it was mentioned above, we study three CVIs-based label noise filtering methods and compare their performance against that of the LOF method. The identification of mislabeled instances (class outliers) may also be viewed as a classification problem for which the objective of each studied method is to distinguish positives (class outlier) from negatives. A true classification is achieved when a method classifies an instance correctly, otherwise the classification is false.

In view of the above, we may benchmark each studied CVIs-based method against LOF as follows: The null hypothesis of interest is that the difference in performance between the studied method and LOF is zero. The test of this hypothesis involves the comparison of two methods on multiple data sets. A suitable method to apply when testing the null hypothesis is a paired t-test (e.g. see Demzar (2006)), which checks whether the average difference in their performance over the data sets is significantly different from zero. In our scenario, we use the accuracy scores generated by the five learning algorithms on the ten data sets in order to evaluate the performance of each filtering method. These scores are used to calculate the paired t-test.

4.2 Implementation and Availability

The discussed CVI-based Outlier Filtering algorithm has been implemented in Python 3.6. In our experiments we have used three different cluster validation measures: Silhouette Index, Connectivity and IC-av. Silhouette Index is used from the Python library Scikit-learn. IC-av index has been implemented in Python according to the description given in Section 3 (see Equation 6) while Connectivity has been coded following its R script definition. We compare the CVI-based outlier filtering algorithm against LOF (Breunig et al, 2000). We have used the implementation of LOF provided by scikit-learn with the default neighborhood size of 20, an optimal number suggested by the authors. In addition, the used neighborhood size of Connectivity is 10, a default value of its R implementation. We have used the *F-measure* to evaluate the accuracy of the learning algorithms used in our experiments. The scikit-learn implementation of the F-measure (*micro-average* F_1) has been used. The executable of the CVI-based Outlier Filtering algorithm, the used data sets and the experimental results are available on GitLab (`https://gitlab.com/machine_learning_vm/outliers`).

4.3 Results and Discussion

Figure 3 depicts the average improvement over all data sets and all learning algorithms for the four studied outlier filtering methods by comparing local (left) versus global (right) filtering. As one can see the best results are achieved by SI $\vee$ Co $\vee$ IC-av with 4 % local filtering and Median(SI, Co, IC-av) with 10 % global filtering, respectively. In addition, SI $\vee$ Co $\vee$ IC-av outperforms LOF for all studied cases. Median(SI, Co, IC-av) is also better than LOF for almost all cases except only for 2 % local filtering. In addition, Median(SI, Co, IC-av) outperforms SI $\vee$ Co $\vee$ IC-av in 3 from totally 10 conducted filtering experiments (10 % local and 8 % and 10 % global). However, LOF performs better than SI $\wedge$ Co $\wedge$ IC-av in all studied filtering scenarios. It is interesting to notice that in the case of global filtering SI $\wedge$ Co $\wedge$ IC-av is getting closer to LOF with the increase of the percentage of filtered instances. This may be due to the fact that SI $\wedge$ Co $\wedge$ IC-av enables to identify more mislabeled instances then LOF when filtering out a higher percentage from the data set.

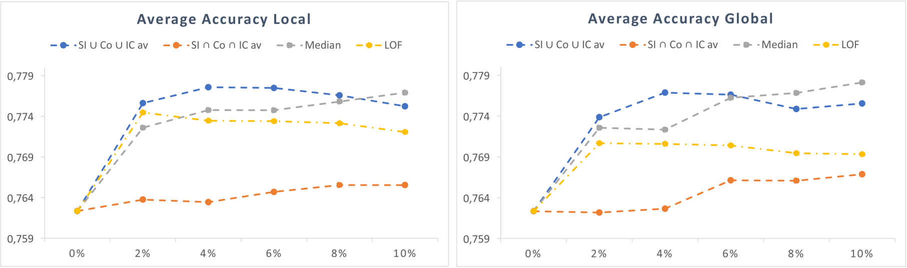

Figure 3: Average gain of classification accuracy over all data sets and all learning algorithms: local versus global filtering.

In order to improve the understandability of the results in the following figures we present the most significant examples from the data sets and learning algorithms. All the other results generated on the different data sets by using the selected learning algorithms can be found on GitLab (`https://gitlab.com/machine_learning_vm/outliers`).

In Figure 4 the results generated by 1-NN learning algorithm on the well-known *Iris* data set are given. In this context the three CVIs-based methods outperform LOF almost in all studied filtering scenarios. Moreover, SI $\vee$ Co $\vee$ IC-av is again the best performing method. Median(SI, Co, IC-av) performs very close to SI $\vee$ Co $\vee$ IC-av in all conducted experiments.

We can see in Figure 5 that the accuracy scores in case of the union of the three CVIs (SI $\vee$ Co $\vee$ IC-av) for filtering with 4 % is quite high in comparison with the other results. The performance slightly degrades with higher percentage of filtering, but it still outperforms LOF. Notice that the three CVIs-based label noise filtering methods outperform LOF in all studied filtering scenarios in this context (GNB conducted on the *Ecoli* data set). It is also interesting to observe that both Median(SI, Co, IC-av) and SI $\wedge$ Co $\wedge$ IC-av in contrast to SI $\vee$ Co $\vee$ IC-av show increased accuracy results with higher percentage of filtering. In general, Median(SI, Co, IC-av) has demonstrated more stable behaviour than the other two CVIs-based filtering methods in the conducted experiments.

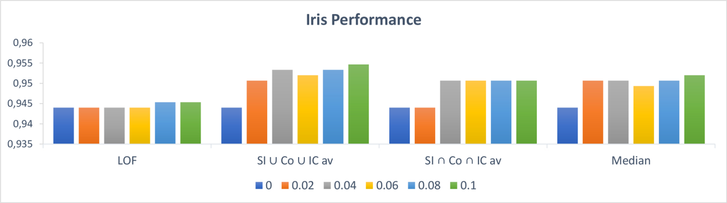

Figure 4: Results for *Iris* with 1-NN evaluation.

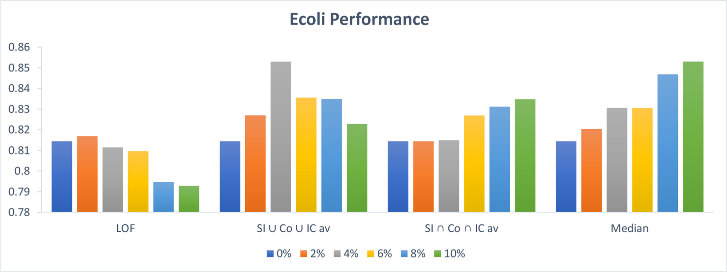

Figure 5: Results for *Ecoli* with GNB evaluation.

It is interesting to mention that for the CART algorithm we have seen the highest quality improvements thanks to the outlier removal. For example, in case of 0 % filtering we have achieved an accuracy of only 0.13. As soon as we start filtering the quality goes up to around 0.50. The latter shows how extreme the impact of mislabeled instances can be on the performance of the classification tasks.

The aforementioned results are confirmed by paired t-tests. The LOF detection method has been benchmarked against each studied CVIs-based label noise filtering method by conducting the paired t-test on the accuracy scores generated by the five learning algorithms on the ten studied data sets with 6 % globally filtered out by the corresponding methods. This filtering percentage is selected since all the four studied methods have shown good performance with it. Hypothesis testing is conducted at $p < 0.05$ and 49 ((5 learning algorithms × 10 data sets) - 1) degrees of freedom. The corresponding critical value for t-test in our case is 2.01. The null hypothesis can be rejected for Median(SI, Co, IC-av) and SI $\vee$ Co $\vee$ IC-av, since the corresponding calculated t-test scores are 4.61 and 2.85. However, the t-test score generated by SI $\wedge$ Co $\wedge$ IC-av is -0.24, i.e. it is below the critical value. The obtained results are also supported by the ones presented in Figure 3.

5 Conclusion and Future Work

In this work, we have proposed an outlier filtering approach, entitled CVI-based Outlier Filtering, that applies cluster validation measures to identify mislabeled instances. The implemented version of the CVI-based algorithm uses three internal cluster validation measures. In addition, we have studied three different assembling strategies (union, intersection and ranked-based median), i.e. three different CVIs-based label noise filtering methods. These have been evaluated and compared against the LOF detection method for five commonly used learning algorithms on ten data sets from the UCI data repository. The obtained results have demonstrated that the proposed approach is a robust outlier filtering technique that is able to improve classification accuracy of the learning algorithms. Our approach allows to build a label noise detecting measure that is specially suited for the machine learning task under consideration. Namely, we can initially study and select a proper combination of cluster validation measures that reflects the specifics of the involved data and learning algorithms.

For future work, the aim is to pursue further enhancement and validation of the proposed outlier filtering approach by applying alternative cluster validation measures on a higher variety of data sets and learning algorithms. In addition, we also plan to study different ranked-based ensemble methods for assembling the selected cluster validation indexes, as well as scoring-based methods. The latter ones raise additional challenges for developing suitable cluster validation indices normalization techniques.

References

Aggarwal CC (2013) Outlier Ensembles: Position paper. ACM SIGKDD Explorations Newsletter 14(2):49–58. DOI: 10.1145/2481244.2481252.

Bayá AE, Granitto PM (2013) How Many Clusters: A Validation Index for Arbitrary-shaped Clusters. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10(2):401–414. DOI: 10.1109/TCBB.2013.32.

Bezdek J, Pal N (1998) Some New Indexes of Cluster Validity. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 28(3):301–315. DOI: 10.1109/3477.678624.

Boeva V, Lundberg L, Angelova M, Kohstall J (2018) Cluster Validation Measures for Label Noise Filtering. In: 9th IEEE International Conference on Intelligent Systems (IS'18), Jardim-Gonçalves R, Mendonça JP, Jotsov V, Marques M, Martins J, Bierwolf R (eds), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 109–116. DOI: 10.1109/IS.2018.8710495.

Breunig M, Kriegel HP, Ng R, Sander J (2000) LOF: Identifying Density-based Local Outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00), Chen W, Naughton JF, Bernstein PA (eds), Association for Computing Machinery (ACM), Vol. 29, pp. 93–104. DOI: 10.1145/342009.335388.

Brodley CE, Friedl MA (1999) Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research 11:131–167. DOI: 10.1613/jair.606.

Demzar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7:1–30. URL: `http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf`.

Frénay B, Verleysen M (2014) Classification in the Presence of Label Noise: A Survey. IEEE Transactions on Neural Networks and Learning Systems 25(5):845–869. DOI: 10.1109/TNNLS.2013.2292894.

Gamberger D, Lavrac N, Džeroski S (2000) Noise Detection and Elimination in Data Preprocessing: Experiments in Medical Domains. Applied Artificial Intelligence 14(2):205–223. DOI: 10.1080/088395100117124.

Handl J, Knowles J, Kell D (2005) Computational Cluster Validation in Post-genomic Data Analysis. Bioinformatics 21(15):3201–3212. DOI: 10.1093/bioinformatics/bti517.

He Z, Deng S, Xu X (2002) Outlier Detection Integrating Semantic Knowledge. In: International Conference on Web-Age Information Management (WAIM 2002), Meng X, Su J, Wang Y (eds), Springer, Berlin, Heidelberg (Germany), Lecture Notes in Computer Science, Vol. 2419, pp. 126–131. DOI: 10.1007/3-540-45703-8_12.

He Z, Xu X, Huang J, Deng S (2004) Mining Class Outliers: Concepts, Algorithms and Applications in CRM. Expert Systems with Applications 27(4):681–697. DOI: 10.1016/j.eswa.2004.07.002.

Hewahi N, Saad M (2007) Class Outliers Mining: Distance-based Approach. International Journal of Intelligent Systems and Technologies 2(1):2792–2805.

Jain A, Dubes R (1988) Algorithms for Clustering Data. Prentice-Hall, Upper Saddle River (USA). ISBN: 978-0-130222-78-7.

Jaskowiak PA, Moulavi F D., C.S. A, Campello RJ, Zimek A, Sander J (2016) On Strategies for Building Effective Ensembles of Relative Clustering Validity Criteria. Knowledge and Information Systems 47(2):329–354. DOI: 10.1007/s10115-015-0851-6.

Khoshgoftaar TM, Seliya N, Gao K (2004) Rule-based Noise Detection for Software Measurement Data. In: Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, pp. 302–307.

Kohavi R (1995) A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95), Morgan Kaufmann Publishers Inc., San Francisco (USA), Vol. 14, pp. 1137–1145. ISBN: 978-1-558603-63-9.

Kubica JM, Moore A (2003) Probabilistic Noise Identification and Data Cleaning. In: 3rd IEEE International Conference on Data Mining (ICDM 2003), Wu X, Tuzhilin A, Shavlik J (eds), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 131–138. DOI: 10.1109/ICDM.2003.1250912.

Papadimitriou S, Faloutsos C (2003) Cross-outlier Detection. In: International Symposium on Spatial and Temporal Databases (SSTD 2003): Advances in Spatial and Temporal Databases, Hadzilacos T, Manolopoulos Y, Roddick J, Theodoridis Y (eds), Springer, Berlin, Heidelberg (Germany), Lecture Notes in Computer Science, Vol. 2750, pp. 199–213. DOI: 10.1007/978-3-540-45072-6_12.

Rousseeuw P (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics 20:53–65. DOI: 10.1016/0377-0427(87)90125-7.

Segata N, Blanzieri E (2010) Fast and Scalable Local Kernel Machines. Journal of Machine Learning Research 11(Jun):1883–1926. URL: `http://www.jmlr.org/papers/v11/segata10a.html`.

Smith M, Martinez T (2011) Improving Classification Accuracy by Identifying and Removing Instances That Should Be Misclassified. In: The 2011 International Joint Conference on Neural Networks (IJCNN), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 2690–2697. DOI: 10.1109/IJCNN.2011.6033571.

Smith M, Martinez T (2016) A Comparative Evaluation of Curriculum Learning with Filtering and Boosting in Supervised Classification Problems. Computational Intelligence 32(2):167–195. DOI: 10.1111/coin.12047.

Tomek I (1976) An Experiment with the Edited Nearest-neighbor Rule. IEEE Transactions on Systems, Man, and Cybernetics SMC-6(6):448–452. DOI: 10.1109/TSMC.1976.4309523.

Vendramin L, Jaskowiak P, Campello R (2013) On the Combination of Relative Clustering Validity Criteria. In: Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM'13), Szalay A, Budavári T, Balazinska M, Meliou A, Saçan A (eds), Association for Computing Machinery (ACM), New York (USA), pp. 733–744. DOI: 10.1145/2484838.2484844.

Zimek A, Schubert E, Kriegel HP (2012) A Survey on Unsupervised Outlier Detection in High-dimensional Numerical Data. Statistical Analysis and Data Mining 5(5):363–387. DOI: 10.1002/sam.11161.