Irregularity Detection in Categorized Document Corpora

Borut Sluban [(1)], Senja Pollak [(2)], Roel Coesemans [(3)], Nada Lavrač [(1,4)]
(1) Jožef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia
(2) University of Ljubljana, Faculty of Arts, Department of Translation, Ljubljana, Slovenia
(3) University of Antwerp, IPrA Research Center, Antwerp, Belgium
(4) University of Nova Gorica, Nova Gorica, Slovenia
E-mail: borut.sluban@ijs.si, senja.pollak@ff.uni-lj.si

Abstract

The paper presents an approach to extract irregularities in document corpora, where the documents originate from different sources and the analyst's interest is to find documents which are atypical for the given source. The main contribution of the paper is a voting-based approach to irregularity detection and its evaluation on a collection of newspaper articles from two sources: Western (UK and US) and local (Kenyan) media. The evaluation of a domain expert proves that the method is very effective in uncovering interesting irregularities in categorized document corpora.

Keywords: text mining, text categorization, irregularity detection

1. Introduction

The process of document corpora analysis which employs text mining methods (Feldman and Sanger, 2007) consists of the following steps: selection of the target corpora, document preprocessing, text mining, and finally the interpretation and human evaluation of the automatically extracted insights. This work considers the problem of detecting and analyzing atypical and/or irregular documents in categorized document corpora, to be used for data cleaning and evaluation of language resources.

Handling erroneous and/or atypical instance in data collections has been studied in various data mining areas. Errors and other irregularities in data are commonly referred to as noise. Detecting and/or eliminating noise is aimed at reducing the negative effects of noise on classification accuracy, knowledge extraction, results interpretation, computational complexity, etc. (Zhu and Wu 2004).

Constructing noise resistant machine learning algorithms by generalizing classification models were the first approaches to noise handling (Quinlan, 1987; Niblett and Bratko, 1987; Fürnkranz, 1997). For labeled data predictions of classification algorithms are used to distinguish between regular and irregular data instances in noise elimination by filtering (Brodley and Friedl, 1999). Using different level of agreement among multiple classifiers was explored by Verbaten and Van Assche (2003) and Khoshgoftaat et al. (2005). The saturation filtering approach presented by Gamberger and Lavrač (1997) identifies noisy instances as those that reduce the complexity of a data representation model.

Atypical or irregular data instance that are not necessarily erroneous on their own are also referred to as outliers[1]. Distribution-based approaches (Van Hulse et al, 2007) and clustering or distance based approaches (Knorr and Ng, 1998; Yin et al., 2009) are widely use for the detection of atypical/outlier instances in unlabeled data. An extensive overview of outlier detection methodologies is presented by Hodge and Austin (2004).

For our task of detecting atypical or irregular documents in categorized document corpora we chose a voting-based irregularity detection approach in which we use different classification algorithms, to identify 'misclassified' or 'falsely categorized' documents, and a saturation filtering approach, that uses the complexity of a document classification/categorization model as a measure for irregularity identification.

This work proposes a method for detecting atypical and/or irregular documents in categorized document corpora and presents a domain expert's evaluation of the detected atypical documents. The analysis of most significant atypical documents was performed by the expert in pragmatics discourse analysis who was able to inspect the documents and explain the reasons underlying their irregularity status. In his opinion, the approach proved to be successful. As this paper shows, the analysis indeed proves the utility of the proposed method for qualitative analysis of document corpora collected from different sources.

This paper is structured as follows. We describe the preprocessing of text documents in Section 2 and the methodology for irregularity detection in Section 3. The experimental settings and the irregularity detection results are presented in Section 4. The linguistics expert evaluation of irregularity detection results on the newspaper articles document collection follows in Section 5. We compare our irregularity detection results to a baseline method in Section 6. The summary and discussion conclude our paper in Section 7.

[1] In statistics, an outlier is an observation that is numerically distant from the rest of the data, or more formally, it is an observation that lies outside the overall pattern of a distribution.

2. Document Preprocessing

After the identification of the target corpus, the documents in the corpus must be preprocessed and converted to the format required by text mining methods. Standard document preprocessing (Feldman and Sanger 2007) is performed as follows:

- Text tokenization: a continuous character sequence is split into meaningful sub-tokens, i.e., individual words or terms.
- Stopword removal: stopwords are predefined words from a language that usually carry no relevant information (e.g. and, or, a, an, the, ...).
- Stemming/Lemmatization: the process that converts each word/token into its morphologically neutral form.
- *N*-grams construction: *N*-grams are terms defined as a concatenation of 1 to *N* words which appear consecutively in the text.
- Bag-of-words (BoW) representation: a vector representation of a document, with value 1 (or word frequency-based weight) for words/terms appearing in the document, and value 0 for the rest of the corpus vocabulary.
- Feature selection: selection of most informative features (words/terms from the BoW representation)[2].

Documents presented in the BoW vector space format can now be processed by text mining methods, such as our irregularity detection method.

3. Irregularity Detection Method

The idea behind irregularity detection in categorized document corpora is based on early noise filtering approaches presented by Brodley and Friedl (1999), who used a classifier as a tool for detecting noisy instances in data, and on a different saturation-based noise filtering approach presented by Gamberger and Lavrač (1997), which identifies noisy instances that reduce the complexity of a data representation model.

Noise detection approaches identify irregularities and errors in data, therefore they are suitable also for detecting atypical, unusual and/or irregular documents in categorized document corpora. Since our aim is to detect atypical documents to be inspected by human experts in the phase of data cleaning and data understanding, we want our approach to identify documents for which one can claim—with high certainty—that they are indeed atypical and are worthy the expert's inspection time and effort.

To obtain more reliable results when identifying atypical documents, prediction of several different noise detection approaches can be taken into account in order to increase the reliability. We use this principle in our voting-based approach to irregularity detection in categorized document corpora, which implements classification- and saturation-based noise detection methods.

- In classification-based noise detection, first proposed by (Brodley and Friedl, 1999), a data instance is denoted as noisy if it is incorrectly classified by one or more classifiers. A cross-validation approach is used: a dataset is partitioned into *n* subsets and repeatedly *n*-1 subsets are used for classifier learning and the complementary subset for evaluating the classifiers. These methods are referred to as *Classification Filters*. In this work, the simple classifiers used in (Brodley and Friedl, 1999) were replaced by better performing classifiers. The following five different classifiers are used in our implementation of the irregularity detection method: the Naïve Bayes classifier (as the baseline filtering algorithm), Random forest classifier with 100 decision trees (RF100), Random forest classifier with 500 decision trees (RF500), Support vector machine classifier (SVM), and Support vector machine classifier with automatic data scaling and parameter optimization (SVMEasy)[3].
- A substantially different noise detection method, the so-called *Saturation Filter* (SatFilt), developed by Gamberger and Lavrač (1997), is based on the observation that the elimination of noisy examples reduces the so-called Complexity of the Least Complex correct Hypothesis (*CLCH*) value of the dataset. The proposed *CLCH* measure is used to find a saturated dataset enabling the induction of a hypothesis that correctly captures the generally valid concept of the domain presented by the available data; noisy examples are those which are outside of the saturated dataset. In our implementation of the saturation filter we use the number of nodes of a decision tree as the complexity measure which corresponds to the *CLCH* value. For more details on the implementation refer to Sluban et al. (2009).

Extensive performance evaluation of the above algorithms for irregularity (outlier, noise) detection was performed in our previous work (Sluban et al. 2011) on four different domains with various amounts of artificially added noise. That paper not only evaluates the individual classifiers but also shows the influence of their combination for the task of irregularity detection.

In this work we used all these algorithms to construct a group of voters consisting of three types of classification noise filters using: Naïve Bayes, Random Forest (two variants: RF100 and RF500) and SVM classifiers (two variants: SVM and SVMEasy), as well as the Saturation Filter (SatFilt). Our voting approach to irregularity detection returns a set of irregular documents, which are

[2] In our case, based on the χ^2-distribution statistic, 500 best features were selected in the BoW vectors used by our irregularity detection method.

[3] We used the implementations of these classification algorithms, which are well known in the data mining and text mining community, from the Orange data mining framework (Curk et al., 2005)

grouped and sorted according to the number of noise detection algorithms that identified a document as noisy. This voting-based approach for irregularity detection enables the domain expert (analyst) to inspect documents grouped by their significance of being atypical or irregular documents of the examined document corpus.

4. Experiments

In order to test our voting-based irregularity detection method for obtaining interesting documents which are atypical for a document collection, we built a corpus of documents, originating from two different newspaper sources. We selected a subset of articles from a larger corpus of newspaper articles originally collected by the IPrA Research Center, University of Antwerp, as part of the Intertextuality and Flows of Information project in which an ethnographically-supported pragmatic analysis of news discourse was undertaken.

In our experiments we analyzed 464 articles (about 320,000 words) originating from six different daily newspapers in English, covering the time period from December 22, 2007 to February 29, 2008, concerning Kenyan presidential and parliamentary elections, held on December 27, 2007, and the crisis following the elections. Articles from the British and US press (The Independent, The Times, The New York Times and The Washington Post) formed the category "Western" (WE) and articles from Kenyan newspapers Daily Nation and The Standard were categorized as "local" (LO). More details on this document collection can be found in Pollak (2009) and Pollak et al. (2011).

We tested our voting-based irregularity detection method on the selected newspaper articles collection using six different noise detection algorithms described in the previous section. We obtained a set of atypical and/or irregular articles grouped and sorted according to the number of noise detection algorithms that identified them as irregular.

Article (Class and ID)	Number of Votes
WE 352	6
LO 25 LO 101 LO 173 WE 348 WE 326 WE 357 WE 410	5
LO 4 LO 21 LO 68 LO 162 WE 358 WE 464	4

Table 1: Articles that were identified as atypical by the majority of noise detection algorithms.

Since our aim is to obtain reliable results and investigate actual/genuine irregularities, we considered only articles that were identified as irregular by the majority of the applied noise detection algorithms (i.e., 4 or more). We list these atypical articles identified by our voting-based irregularity detection method in Table 1.

5. Results Interpretation and Evaluation

Detailed analysis of the top 14 atypical articles identified by our voting-based irregularity detection method was performed by the domain expert from the field of linguistic pragmatics. Instead of explaining the detected atypical documents one by one in the order as grouped together by their number of votes in Section 4, we have – for the sake of this presentation – grouped the articles by the nature of their irregularity type. Irregularity group A contains articles that in fact turned out to be problematic and should not have been incorporated in the document corpus: based on this analysis we will propose to the linguists who originally collected the corpus to exclude these articles from the corpus. The articles of Irregularity group B are the articles that best prove the effectiveness of the irregularity detection method. In our case we have three articles published in local press, being written by 'Western' journalists or talking about 'Western press'. In Irregularity group C there are articles which are specific in their genre, being more opinion type of articles rather than 'hard news' type of articles. The Irregularity group D contains three articles that are special because of their extreme document length. One of the detected articles, however, did not belong to any of the uniform irregularity types.

5.1 Irregularity group A: Out of topic

Several articles were detected as anomalies because their content is not exactly within the topic which is central to our analysis. When building a corpus for qualitative linguistic analysis the articles were collected using the keywords related to the main topic of interest. Irregularity detection can be viewed as a very effective way for quickly detecting documents which were included based on the keywords and are in fact out of the main topic of interest: they actually represent mistakes made in document corpus collection.

Article **WE 352** which was voted as most atypical (as it was misclassified by all the classifiers and declared as noisy also by the Saturation filter) is out of topic because it is not about the Kenyan elections or post-election violence but is about violence in Kenya. It can be explained from the following point of view. When the corpus of newspaper texts on the specific theme of Kenyan elections and postelection crisis was built, the selection was based on several keywords. In this article the elections are specially mentioned in the content that it is "not connected to postelection violence". It contains the key term but the article describes a criminal attack (a robbery for money), which is neither by the journalist, nor by its context related to the postelection violence. The

article was later indeed removed from the corpus used for further linguistic analysis, since it is not about the socio-political climate but about British tourists or expatriates' misfortune.

Another reason for its misclassification might be that it reports on violence in Kenya without making a link to ethnicity, rather focusing on socio-economic motives. This is typical of the Kenyan coverage of the election crisis. While the Western newspapers usually mentioned the ethnicity of perpetrators of violence, the Kenyan newspaper shied away from ethnic labels. Here the ethnicity of the robbers is not mentioned. In Kenyan newspaper references to people involved in violence are often vague, unspecified or general, comparable to this article's reference to "a gang of six men", as well as the use of the term *community* instead of *tribe*. Note that in this article the typical 'local-class word' *community* is used, but it has a different meaning.

Also article **WE 348** is not about the central topic. It provides travel information and highlights a different aspect of the crisis in Kenya: declining number of tourists, while tourism is an important economic sector in Kenya. This contrasts with Kenyan press coverage where the economic aspects of the crisis are more frequently discussed.

Article **WE 464** only briefly mentions the crisis in Kenya, but mainly deals with U.S. foreign policy. As such, containing main keywords, but not covering the Kenyan elections topic, it can be considered as an error in data collection.

5.2 Irregularity group B: Western journalists

This is a very interesting group of irregular articles. We mention here three local articles which are of great interest. First two are interesting because their author is in fact not a Kenyan journalist, therefore confirming the real 'irregular nature' of the article, while the third article's topic justifies its 'irregular' nature.

Article **LO 173** was identified as irregular, as it can be regarded as being a 'Western article' among the local Kenyan articles category. It is written by a Canadian freelance journalist and travel writer, who at the time worked for the Daily Nation. It was also observed this journalist does not have the cultural sensitivity or does not follow the editorial guidelines requiring the journalists to be careful when mentioning words like *tribe* in negative contexts. The journalist has a kind of 'Western' writing style. Although this article is published in the section of national news, traditionally covering factual news, this article also describes a topic of general human interest (mixed marriages).

Also article **LO 4** was not written by a Kenyan but by a British psychologist who works in Nairobi. Note also that this article, published in the national news section, is not purely factual, but has an opinionating flavor.

The next example is very interesting. Article **LO 162** is not an article by a guest journalist. It is a local Kenyan article about Western news discourse, so it contains a lot of Western voices. It consists of quotes from the international newspapers, such as Financial Times, The Guardian and The Daily Telegraph. It discusses the Western news reporting of the Kenyan election crisis.

5.3 Irregularity group C: Different genre

Most articles in the corpus are 'hard news' reports, i.e. reports on actual events, but there are also some editorials and opinion articles present. Several of these articles were detected as irregularities.

Article **LO 101** is an opinion article taking a historical perspective and a macro-perspective. This is not typical of the Kenyan press coverage, which tends to focus on situated, localized incidents rather than putting the events in the macro-frame of nation, history, economics, or - as Western newspapers often do - ethnicity. The Kenyan press did not openly characterize the whole elections as rigged (although they admit rigging in certain districts or constituencies). But in this article the word "rigged" is prevalent possibly because the author makes a historical comparison with Uganda and the rigging is especially related to Uganda. On the one hand, this article is a bit off the main topic, as it has a large part about Ugandan political history. Moreover, the broad African perspective and comparisons to other African countries (e.g. Rwanda) occurred much more in the Western press than in the local press.

Also article **LO 25** is a different subgenre than 'hard news'. It is a personal portrait of the president and a hypothetical analysis of what might happen in some likely election scenarios. The emphasis on personal characteristics is in general more typical of the Western press because of different backgrounds of the readership. Also the mentioned stereotypes about the man (e.g. him being a gentleman of Kenyan politics, a technocratic economist with no feeling for the common man, etc.) feature prominently in the Western press, but rarely in the Kenyan press.

Article **LO 21** is a very biased opinion article (appearing in the Opinion and Analysis section), almost political propaganda. It gives a preview of the upcoming elections. Also article **LO 68** is an opinion article.

5.4 Irregularity group D: Extreme document length

Two articles (**WE 357**, **WE 358**) are extremely short, each containing just two sentences each and appearing in short notices section. On the other hand, one document (**WE 410**) was included in the corpus by mistake and was in fact a batch of numerous different articles from different sources, comprising more than 400 pages, due to a data collection error.

5.5 Uncategorized irregularity

For one article, **WE 326**, we are uncertain, whether it is an irregularity or not. Nevertheless, the paper written by a Western author resembles the articles of local journalists in the sense that it covers the parliament, while the Western press focused on the presidency (compared to local press which did deal substantially with parliament news coverage). But there are other Western articles which do report on parliament matters.

6. Comparison to a Baseline Approach

In order to evaluate the performance of our irregularity detection approach we compared the detected atypical documents also by examining how central they are compared to other articles of the same class (LO and WE, respectively). In this setting, an atypical article for a given class is a document which is more similar to documents of the opposite class than to other documents of its own class. The cosine similarity (the cosine of the angle between the vector representations of two documents) is used as a baseline similarity measure. We have modeled the article's irregularity by the difference between the article's similarity to the centroid of the opposite class and the similarity to the centroid of its own class. If this difference is positive the article can be considered as atypical, since this means that it is more similar to the centroid of the opposite class than to the centroid of its own class, and the larger the difference the more atypical is the document for its own class.

To see whether an article is atypical or irregular for the collection of 'local' (LO) and 'Western' (WE) newspaper articles about Kenyan elections and the crisis following the elections, we compared the articles' cosine similarities to the centroid of the LO articles and the centroid of the WE articles. The differences in cosine similarity of an article to the two centroids are presented in Figures 1 and 2 for articles of the LO and WE class, respectively.

In Figures 1 and 2 we labeled 11 documents that were identified as atypical or irregular by our voting-based irregularity detection approach (achieving a majority of votes) as well as by the baseline cosine similarity based irregularity modeling approach. Articles that were not considered irregular by the baseline method, but detected by our voting-based irregularity detection method, were clearly identified by the domain expert as: an off topic article (WE 348), an extreme length article that is too short to be used in further linguistic analysis (WE 357) and a data collection error (WE 464). All other document (except for one) that are considered as irregular according to their difference in cosine similarity to the centroids were also identified by our voting-based irregularity detection method, but did not achieve a majority of votes.

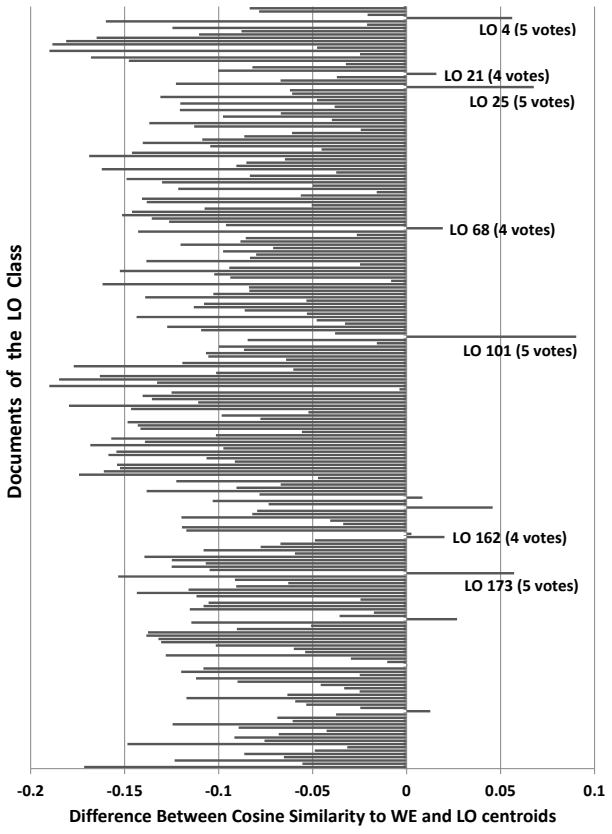

Figure 1: Differences between cosine similarity to the WE centroid and the LO centroid for articles of the LO class.

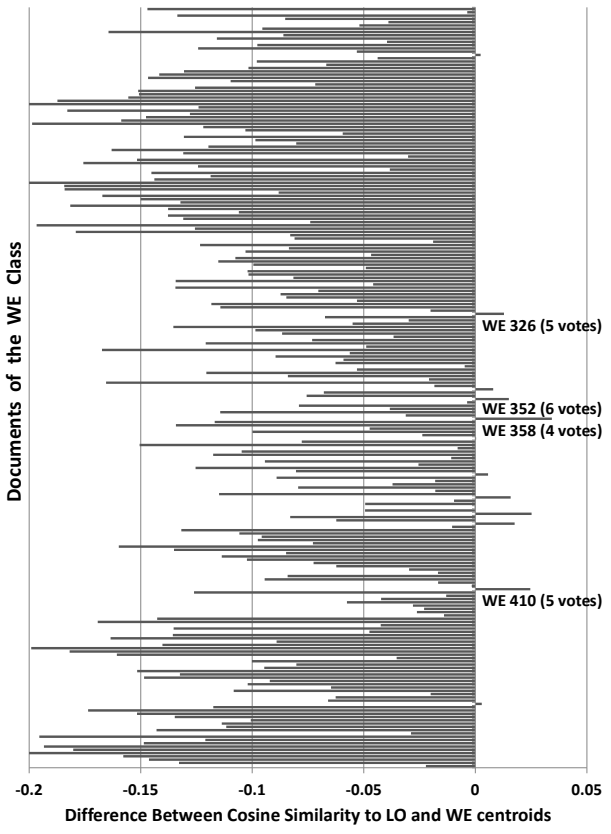

Figure 2: Differences between cosine similarity to the LO centroid and the WE centroid for articles of the WE class.

7. Summary and Discussion

This work presents a voting-based irregularity detection method which can be successfully applied in the process of data cleaning and data understanding of categorized document corpora.

This paper is limited to the application of our method on a single domain of articles from two different classes/categories of newspapers, and the results evaluation by a domain expert. We are aware of the limitations of this setting. In expert evaluation at least two expert opinions would need to be compared in order to compute an inter-annotation agreement. However, as in this work we were mainly interested in the pragmatics discourse analysis aspects of this interpretation, inter-annotation agreement is left for further work. Furthermore, for the qualitative evaluation of the performance of the proposed approach, a study on more than one domain would have been needed and is planned for further work. Quantitative performance evaluation on more domains was addressed in our previous work (Sluban et al., 2011) where our voting-based irregularity detection approach was evaluated on four domains with various levels of artificially inserted noise.

The above mentioned incompleteness of the evaluation setting used in this work is compensated by confirming that the detected irregular documents are indeed atypical for their own class, as they are more similar to documents of the other class than to documents of their own class. Moreover, the expert evaluation of irregular articles shows that the proposed voting-based irregularity detection approach is very effective in irregularity detection and can help the domain expert to discover various types of irregularities in the data. This enables the expert to better understand the data and thereby supports his or her further decision making and further actions in the data analysis process.

8. Acknowledgements

This work has been partially funded by the European Commission in the context of the FP7 project FIRST, Large scale information extraction and integration infrastructure for supporting financial decision making, under the grant agreement n. 257928.

9. References

Brodley, C.E., Friedl, M.A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11, pp. 131-167.

Curk, T., Demšar, J., Xu, Q., Leban, G., Petrovič, U., Bratko, I., Shaulsky, G., Zupan, B. (2005). Microarray data mining with visual programming. *Bioinformatics*. 21(3), pp. 396-398.

Feldman, R., Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data*. Cambridge University Press.

Gamberger, D., Lavrač, N. (1997). Conditions for Occam's razor applicability and noise elimination. *In Lecture Notes in Artificial Intelligence: Machine Learning: ECML-97*, 1224, pp. 108-123.

Hodge, V., Austin, J. (2004). A survey of outlier detection methodologies, *Artificial Intelligence Review*, 22(2), pp. 85-126.

Khoshgoftaar, T. M., Zhong, S., Joshi, V. (2005). Noise elimination with ensemble classification filtering for software quality estimation. *Intelligent Data Analysis*, 9(1), pp. 3-27.

Knorr, E. M., Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets, *In Proceedings of the 24th international conference on very large databases (VLDB)*, pp. 392-403.

Pollak, S. (2009). Text classification of articles on Kenyan elections. In Z. Vetulani (ed.) *Proceedings of the 4th Language & Technology Conference: Human language technologies as a challenge for computer science and linguistics*, pp. 229-233.

Pollak, S., Coesemans, R., Daelemans, W., Lavrač, N. (2011). Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining. In *Pragmatics: Quarterly Publication of the International Pragmatics Association*, 21(4), pp. 647-683.

Sluban, B., Gamberger, D., Lavrač, N., Bauer, A. (2009). Experiments with saturation filtering for noise elimination from labeled data. In *Proceedings of the 12th International Multi-conference Information Society - IS 2009*, volume A, pp. 240-243.

Sluban, B., Gamberger, D., Lavrač, N. (2011). Performance analysis of class noise detection algorithms. In: T. Ågotnes (ed.) *STAIRS 2010 - Proceedings of the Fifth Starting AI Researchers' Symposium*, pp. 303-314.

Verbaeten, S., Van Assche, A. (2003). Ensemble methods for noise elimination in classification problems. In: T. Windeatt, F. Roli (eds.) *Multiple Classifier Systems, Lecture Notes in Computer Science*, 2709, pp. 317-325.

Yin, H., Dong, H., Li, Y. (2009), A cluster-based noise detection algorithm. *International Workshop on Database Technology and Applications*, 0, pp. 386-389.

Zhu, X., Wu, X. (2004). Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts, *Artificial Intelligence Review*, 22, pp. 177-210.