International Journal of Computational Linguistics and Applications, Vol. 6, No. 2, 2015, pp. 159–175
Received 16/01/2015, Accepted 27/02/2015, Final 09/09/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

Combining Textual and Visual Features to Identify Anomalous User-generated Content

LUCIA NOCE
IGNAZIO GALLO
ALESSANDRO ZAMBERLETTI
University of Insubria, Italy

ABSTRACT

Anomaly detection has extensive use in a wide variety of applications, such techniques aim to and patterns in data that do not conform to expected behavior. In this work we apply anomaly detection to the task of discovering anomalies from user-generated content of commercial product descriptions. While most of the other works in literature rely exclusively on textual features, we combine those textual descriptors with visual information extracted from the media resources associated with each product description. Given a large corpus of documents, the proposed system infers the key features describing the behavioral traits of expert users, and automatically reports whenever a newly generated description contains suspicious or low quality textual/visual elements. We prove that the joint use of textual and visual features helps in obtaining a robust detection model that can be employed in an enterprise environment to automatically mark suspicious descriptions for further manual inspection.

Keywords: User-generated commercial content, anomaly detection, visual features, textual features, one-class support vector machine

1. INTRODUCTION

Anomaly detection is an important problem that finds interest and usage in many research areas and application domains, e.g. activity

monitoring, fault diagnosis, satellite image analysis, time-series monitoring, pharmaceutical research, medical condition monitoring, detecting novelties in images, detecting unexpected entries in databases or detecting novelty in text [1-3] to name a few.

The main purpose of anomaly detection systems is to identify and, if necessary, remove anomalous observations from data. An anomalous observation, commonly referred to as *outlier*, is defined as a pattern in data that do not conform to a well-defined notion of normal behavior [2].

According to different application domains, outliers often correspond to important and prosecutable information, and arise because of human error, fraudulent behavior, instrument error, natural deviations in populations or faults in systems, *e.g.* an anomalous credit card transaction could detect fraudulent applications for credit cards; anomalous traffic data could correspond to unauthorized access in computer networks; anomalies in magnetic resonance images may suggest presence of malignant tumors [4-6].

In this manuscript we apply anomaly detection to a novel task, casting the problem of discovering fake or suspicious user-generated descriptions of commercial products as an anomaly detection problem; where an outlier description is one that contains textual or visual elements that differ from a notion of canonical behavior inferred from a large corpus of genuine and high quality handcrafted descriptions.

The task of automatically identifying fake or low quality user-generated content is particularly interesting, as nowadays many websites allow users to post their own content (comments, posts, tweets, digital images, video, audio files, *etc.*) to provide a complete and social user experience; while this may increase the credibility of the website and thus increase the user base, it also exposes the platform to a wide number of possible threats, e.g. fake, low quality or malicious user-generated content may damage the credibility of the website and, in some cases, lead to legal sanctions against the website itself.[1]

[1] TripAdvisor fined $600,000 for fake reviews. http://www.cnbc.com/id/102292002

In this work, we inspect whether it is possible to define a system that automatically and effectively marks potentially fake user-generated content coming from a platform that allows users to buy/sell commercial products and to directly provide product descriptions that may include both textual and media information (images and videos).

Unlike other works in literature that focus exclusively on analyzing the textual information [7-11], in our work we also exploit visual attributes extracted from the images attached to the user-generated content, building an innovative and more complete set of features that better describes the typical characteristics that a canonical high quality product description should possess. Such set of textual and visual features is used to train a machine learning model [12], achieving excellent detection rates and fast recognition times.

2. RELATED WORKS

The topic of this work is strongly related to the following research areas: Anomaly Detection, Image Analysis and One Class Classification.

In literature, anomaly detection is primarily applied on text data as *novelty detection*, with the main aim of detecting novel topics, news stories or events in a collection of documents. Anomalies, or novelties, arise because of a newsworthy event or the presence of a different topic in a particular document, *e.g.* Baker *et al.* [11] address this kind of problem using probabilistic generative models; the more recent work of Blanchard *et al.* [10] approaches novelty detection using semi-supervised models; Mahapatra *et al.* [9] exploit contextual text information to detect anomalies in text data.

In their works, Guthrie *et al.* [7, 8] specialize on finding outliers in documents considering several aspects such as topic, author, genre or emotional tone, and use a combination of stylistic and lexical features to detect anomalies. Their unsupervised model can reliably identify outliers composed of 1000 or more words [8]; a paragraph of text is classified as

anomalous when it is irregular, or when it substantially deviates from its surroundings.

Several anomaly detection techniques have also been applied to images, *e.g.* satellite imagery, spectroscopy, mammographic image analysis and video surveillance [6, 13-15]. Outliers in images are usually identified as either single anomalous points/pixels or as entire sub-regions, and they are detected by analyzing several image attributes such as color, lightness and texture.

In our work we combine the two previously cited classes of anomaly detection methods (text and images), to find outliers on the basis of both the quality of the textual information provided by users and the overall quality of the images associated with those textual elements.

Assessing the overall quality of natural images is a difficult task, as many image attributes need to be taken into account and the notion of "quality" is often subjective. One of the simplest and more relevant image quality measure is provided by focus measure operators [16], which typically analyze the spectrum of the given natural image to detect the presence of strong edges and determine whether the image has been acquired at the right distance from the camera sensor. Another image quality measure that we consider relevant to determine fake and low quality user-generated product descriptions is the presence of hyperlinks within images attached to the textual content, as those hyperlinks usually lead to malicious third party websites.

Since it is difficult and time expensive to collect a decent amount of fake, low quality or badly written user-generated product description, our method is closely related to all those One Class Classification works that try to classify positive cases without exploiting information collected from negative ones [17].

Many works from One Class Classification research area are focused on text classification, e.g. Liu et al. [18] treat the problem of building text classifiers using positive and unlabeled examples, using a biased formulation of Support Vector Machine (SVM) [12], obtaining state-of-the-art results; Manevitz and Yousef [19] compare different One Class Classification models

in the context of Information Retrieval, showing that SVM and Neural Networks obtain the highest detection rates.

In our work, we exploit information extracted from high quality user-generated product descriptions to train a one class SVM classifier that detects anomalous user-generated product descriptions by jointly analyzing different type of data.

3. PROPOSED METHOD

The proposed approach is presented in this section: we define a set of features describing the goodness of both textual descriptions (Section 3.1) and related images (Section 3.2), and those textual and visual features are used to train a one class SVM model (Section 3.3).

3.1. *Textual features*

We analyze the textual information associated with each product description focusing on the aspects that, in our application context, usually characterize a high quality user-generated product description. In the following paragraphs, we illustrate the aspects we have taken into account, pointing out the intuitions behind each of them.

Description Length. When focusing on finding abnormal/ anomalous product descriptions, one of the first feature that needs to be taken into account is the length of the description in terms of number of text characters. The assumption behind this feature is that both an excessively long or short product description should rise a warning. In fact, in our application context, most expert users tend to provide descriptions that have similar lengths, and thus the variance in terms of number of characters between different high quality product descriptions is usually small. Using this numerical feature, our machine learning model learns the distribution of lengths from the high quality product descriptions in the positive training set, and spots possible outliers during the generalization phase.

Description Language. In our application context, the language used to write the product description is particularly relevant, as high quality product descriptions should always be written using the platform main idiom. To identify the language used in a description we rely on the Language Detection Library for Java [20], which not only detects the main language used in the processed document, but also provides a list of percentages of use of all the other languages found in the same document.[2] We exploit this feature to describe the language of a document as the percentage of use of the main platform idiom in the document itself; the idea behind this choice is that, although some technical words in a document may belong to different languages (product name, seller's contact information, etc.), the platform main language should have a high usage percentage rate.

Description Keywords. In the dataset we collected, and more generally in online marketplaces, it is extremely common to find, associated with each product description, a set of special words, called *tags*, that typically represent a list of keywords of the description itself. In our work, we use those *tags* to compute an index of consistency of the textual information provided by users. In details, for each document, we measure the percentage of *tag* words that appear in the text description. The intuition behind this consistency index is that fake or low quality product descriptions usually have either no *tags* or random and meaningless *tags*.

Presence of Hyperlinks. The presence of hyperlinks strongly characterizes fake and low quality product descriptions. An external reference or a private email addresses usually redirects the user to a third party competitor/advertisement website and should not be tolerated unless most of the other product description quality measures (length, language, *tags*, image quality, etc.) are consistent with those extracted from high quality genuine product descriptions; in which case it may be possible that the hyperlink simply redirects to a media content, e.g. video,

[2] http://code.google.com/p/language-detection/

that cannot be directly attached to the product description. We detect the presence of both email addresses and website URLs in product descriptions using a set of regex-based rules, and, for each document, we use the number of external hyperlinks in the associated product description as input feature to our one-class SVM model.

(a) focused (b) unfocused

Figure 1. *Examples of focused (a) and unfocused (b) images of commercial products correctly discriminated by Variance of Laplacian [21] focus measure operator.*

3.2. *Visual features*

The images associated with each product description are a valuable source of information and may help in determining the quality and genuineness of the description itself. The main intuition is that, in high quality documents, a product image should clearly shows the object described in the textual description, without reporting further irrelevant or malicious information.

Even though classifying the content of natural images is difficult and expensive in terms of both computational power and time required to extract visual features and train the classification model, it is still possible to define some simpler image quality measures that are relevant to our application context. In particular, for each user-generated image, we take into account both its focusness level, and the presence of hyperlinks within the image itself. Unfocused images are usually associated with low-quality product descriptions, while images containing hyperlinks

leading to third party websites are typically associated with fake or low quality product descriptions.

The two previously cited image quality indexes used in our work are briefly described in the following paragraphs.

Focusness. In order to determine the best focus measure operator for our application context, we compared most of the focus indexes proposed in literature throughout the last decade [16] over focused images extracted from our set of high quality genuine product descriptions. In our experiments, Variance of Laplacian (LAPV) [21] provided the best results both in terms of computational complexity and discriminative power, as it almost always associated high focus values to our positive focused images. Figure 1 shows some examples of focused and unfocused images that LAPV correctly processes (high and low focus values for focused and unfocused images respectively). In our pipeline, for each processed document, the average output of LAPV for all the images associated with the document is provided as input feature to our one-class SVM model. As previously stated, a high average focusness value should denote a finely crafted user-generated product description.

(a) allowed text (b) external references

Figure 2. *Examples of images of commercial products containing allowed text words (a) and malicious hyperlinks/email addresses (b)*

Presence of Visual Hyperlinks. As described in Section 3.1, hyperlinks typically appear in low quality product descriptions, as they are used to redirect users to third party websites. In some cases, malicious users are aware that their product descriptions

may be penalized whenever they contain external references, therefore, instead of injecting hyperlinks into the textual content, they embed them into the images of products. To detect those image hyperlinks, we exploit a properly trained version of the Tesseract Optical Character Recognition Engine [22]. It is important to note that not all the text found in an image should be considered anomalous, e.g. Figure 2 shows several examples of images containing allowed text elements and abnormal or malicious text components. For this reason, once and if text among an image is detected, we apply the same regex-based rules of Section 3.1 to determine whether it represents a hyperlink/email address or not. Equivalently to our *Presence of Hyperlinks* textual feature, for each product image, the number of hyperlinks and email addresses found in the image are used as input feature to our classifier.

3.3. *Proposed model*

As described in Section 2, we approach the problem of finding fake or low quality commercial product descriptions as an anomaly detection problem due to the fact that, in our application context, the amount of available negative data is not large enough to train a binary classifier.

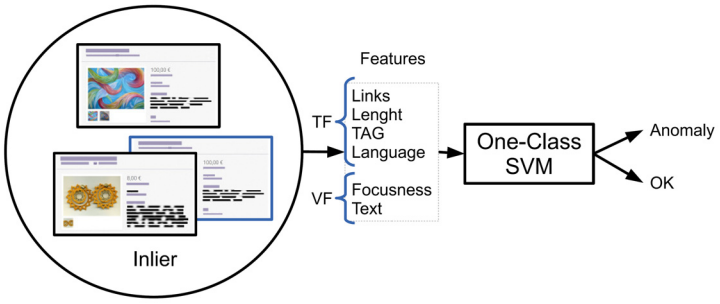


Figure 3. *Visual representation of the training pipeline for the proposed method. The one-class Support Vector Machine model is trained using both textual and visual features (TF and VF respectively) extracted from high quality genuine commercial product descriptions (*inliers*)*

Similarly to other works in literature, we employ a one-class SVM model to infer the behavioral traits of expert users from a dataset composed of high quality genuine product descriptions (described in Section 4.1), using the set of textual and visual features defined in Sections 3.1 and 3.2 respectively. The motivation behind the use of a one-class SVM is that, in anomaly detection, such machine learning model usually lead to optimal results with minimal tuning effort [19].

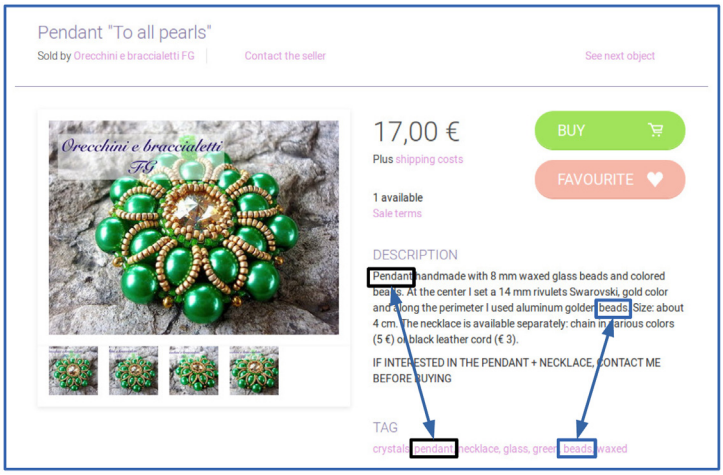
The training phase of the proposed method is summarized in Figure 3. Each positive *inlier* document, typically composed of a textual description and one or more images, is represented as a vector, whose components are the normalized values of the textual TF and visual VF features described in Sections 3.1 and 3.2.

The one-class SVM model is trained using only positive examples, as in classical anomaly detection approaches. When training using a decent amount of significant positive data, we expect the classifer to automatically detect whenever a new product description, that has never been seen during the training phase, contains some anomalous components that substantially differ from those of high quality genuine product descriptions.

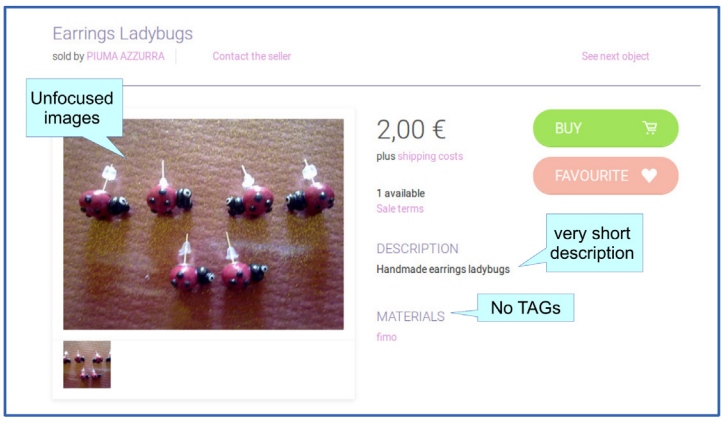
Throughout our experiments (see Section 4), we prove that the proposed set of textual and visual features is exhaustive for the proposed task and that the pipeline described above performs well, achieving high detection rates on the collected dataset.

4. EXPERIMENTS

In this section, we provide an experimental evaluation of the components described in Section 3, and we describe the dataset used in our experiments.



(a) Positive



(b) Negative

Figure 4. *Positive and negative samples extracted from the dataset used in our experiments. (a) A high quality genuine commercial product description containing: an exhaustive textual description,* tags*, a well focused image,* etc. *(b) An artificially generated low quality commercial product description obtained by injecting random anomalies into a high quality user-generated document.*

4.1. *Dataset*

We gather a dataset composed of 41635 documents from a website specialized on selling handmade products. Each insertion is user-generated and composed of a textual description, a set of keywords or tags and one or more images representing the item. All the collected high-quality descriptions documents were manually validated by the administrators of the website.

As in other anomaly detection works [5, 7], the set of possible anomalous descriptions has been artificially built by randomly injecting different kinds of anomalies into high quality genuine user-generated documents, trying to cover a large set of possible anomalous behaviors.

More in detail, we artificially create a total of 1000 negative documents:

- 100 anomalous documents for each type of anomaly that our classier may detect: anomalous length, anomalous language, missing/wrong tags, presence of hyperlinks/email addresses, unfocused product images, and images containing hyperlinks.
- 400 anomalous documents containing two or more types of randomly generated anomalies.

Figure 4 shows a positive high quality product description and an artificially created negative product description containing more than one anomaly. In the positive document of Figure 4 (a), the image of the product is well focused and contains allowed text. On the other hand, in the negative example of Figure 4 (b), the image is unfocused and the item is not clearly displayed. Moreover, the negative example contains a very short textual description and no tags, while in the positive document the object is exhaustively described and two tags are correctly matched in the textual description.

In our experiments, 80% of the high quality genuine documents are used for training, while both the remaining 20% and the artificially created anomalous descriptions are used for testing.

4.2. *Results*

In our experiments, we use the implementation of one-class SVM with RBF kernel provided by the LibSVM Library [12]. The one-class SVM model is trained using 33308 randomly sampled positive documents from the dataset described in Section 4.1. As shown in Figure 3, each document is processed as a vector whose components are the values of the textual and visual features introduced in Sections 3.1 and 3.2 respectively.

As in other works in literature [23, 24], when the number of available positive and negative test documents is strongly unbalanced, it is a common practice to split the set of positive documents into multiple sets having the size of the set of negative documents, and then average the results obtained over each split to compute the final overall result. In our experiments, we divide the 8327 positive test documents into 8 equal splits, and we build 8 different test sets by adding each split to the set of anomalous documents. Each test set is composed of 1040 positive and 1000 anomalous documents.

The goodness of the proposed system at detecting anomalous commercial product descriptions is measured using the following evaluation metrics:

$$\text{Accuracy} = \frac{\text{anomalies correctly identified} + \text{anomalies wrogly identify}}{\text{total \# of documents}}$$

$$\text{Precision} = \frac{\text{anomalies correctly identified}}{\text{total \# of documents marked as anomalous}}$$

$$\text{Recall} = \frac{\text{anomalies correctly identified}}{\text{total \# of anomalies}}$$

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

The overall accuracies obtained by the proposed model are presented in Table 1. They are computed as the arithmetic mean of the accuracies obtained by the trained one-class SVM model on the 8 1040=1000 test datasets. It can be observed that some positive documents are classified as anomalous as their vectorial representations lay on the separation boundary between positive

and negative documents learned by the one-class SVM during the training phase.

Table 1. *Confusion matrix and evaluation results*

	Anomalies	Correct
Identified as anomalous	858	94
Identified as correct	142	946

Accuracy	0.88
Precision	0.90
Recall	0.85
F-measure	0.88

To evaluate the capability of the proposed method at detecting different types of anomalies, for each type of anomalous content considered in our work (length, language, *tags*, hyperlinks in text, focusness and hyperlinks in image) a detection rate has been calculated. For every anomaly type, the respective detection rate is defined as follow:

$$\text{Detection Rate } (\%) = \frac{\text{anomalies correctly identified}}{\text{total } \# \text{ of anomalies}}$$

Results are provided in Table 2. Commercial product descriptions containing more than one type of anomalous behavior are almost always detected, while descriptions that have been injected with single anomalous behaviors are harder to detect, especially the ones containing non matching *tags*. Unsurprisingly, descriptions injected with anomalous visual behaviors (lack of focus and presence of hyperlinks in images) are easily detected, highlighting the importance of considering both visual and textual informations when searching for fake or low quality user-generated content.

Table 2. *Detection rate with different types of anomalies*

Anomalies	Detection Rate
Anomalous length	39%
Anomalous language	79%
Missing/wrong tags	43%
Presence of hyperlinks	99%
Unfocused images	97%
Images containing hyperlinks	98%
$\geq$ 2 random anomalies	97%
Average	78%

In terms of computational complexity, the proposed model requires on average roughly 220 ms to process the textual information associated with each textual description, and a total of roughly 350 ms to process all the visual elements associated with the same description (67 ms for LAPV focus measure operator and 287ms for Tesseract OCR hyperlink extraction).

5. CONCLUSION

A novel anomaly detection method for identifying fake and low quality usergenerated commercial product descriptions has been proposed, it exploits features extracted from both textual and media content to obtain excellent detection rates. Thanks to the use of both a focus measure operator that computes the overall quality level of images, and a text localization/recognition system that identifies hyperlinks and email addresses within images, the proposed system effectively recognizes unusual product descriptions that cannot be detected when exclusively analyzing their textual content. The use of a compact set of highly discriminative features enables our system to process user-generated commercial product descriptions in real time, marking the ones that potentially contain suspicious content for further manual inspection.

REFERENCES

1. Hodge, V. & Austin, J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22, 85-126.

2. Chandola, V., Banerjee, A. & Kumar, V. 2009. Anomaly detection: A survey. *ACM Computing Surveys*, 41, 1-58.
3. Pasha, M. Z. & Umesh, N. 2012. Outlier detection: Applications and techniques. *International Journal of Computer Science*, 9, 307-323.
4. Brause, R., Langsdorf, T. & Hepp, M. 1999. Neural data mining for credit card fraud detection. In *IEEE International Conference on Tools with Artificial Intelligence*.
5. Fan, W., Miller, M., Stolfo, S. J., Lee, W. & Chan, P. K. 2001. Using artificial anomalies to detect unknown and known network intrusions. In *IEEE International Conference on Data Mining.*
6. Spence, C., Parra, L. & Sajda, P. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*.
7. Guthrie, D., Guthrie, L., Allison, B. & Wilks, Y. 2007. Unsupervised anomaly detection. In *International Joint Conferences on Artificial Intelligence*.
8. David Guthrie, L.G. & Wilks, Y. 2008. An unsupervised approach for the detection of outliers in corpora. In *International Conference on Language Resources and Evaluation*.
9. Amogh Mahapatra, N. S. & Srivastava, J. 2012. Contextual anomaly detection in text data. *Algorithms*, 4, 469-489.
10. Blanchard, G., Lee, G. & Scott, C. 2010. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11, 2973-3009.
11. Baker, D., Hofmann, T., McCallum, A., Yang, Y. 1999. A hierarchical probabilistic model for novelty detection in text. In *International Conference on Machine Learning*.
12. Chang, C. C. & Lin, C. J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 1-27.
13. Theiler, J. & Cai, D. M. 2003. Re-sampling approach for anomaly detection in multispectral images. In *International Society for Optical Engineering*.
14. Chen, D., S.X.H.B. & Su, Q. 2005. Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. *Analytical Sciences*, 21, 161-167.
15. Pokrajac, D. 2007. Incremental local outlier detection for data streams. In *IEEE Symposium on Computational Intelligence and Data Mining*.
16. Pertuz, S., Puig, D. & Garcia, M.A. 2013. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46, 1415-1432.

17. Khan, S.S. & Madden, M. G. 2010. A survey of recent trends in one class classification. In *Irish Conference on Artificial Intelligence and Cognitive Science*.
18. Liu, B., Dai, Y., Li, X., Lee, W.S. & Yu, P. S. 2003. Building text classifiers using positive and unlabeled examples. In *International Conference on Data Mining*.
19. Manevitz, L. M. & Yousef, M. 2002. One-class svms for document classification. *Journal of Machine Learning Research*, 2, 139-154.
20. Shuyo, N. 2010. Language detection library for java (2010) Software available at http://code.google.com/p/language-detection/.
21. Pacheco, J. P., Cristobal, G., Martinez, J. C. & Valdivia, J. F. 2000. Diatom auto focusing in brightfield microscopy: A comparative study. In *IEEE International Conference on Pattern Recognition*.
22. Smith, R. 2007. An overview of the tesseract ocr engine. In *International Conference on Document Analysis and Recognition*.
23. Chawla, N. 2005. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 853-867).
24. Bekkar, M., Djemaa, H. K. & Alitouche, T. A. 2013. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3, 27-38.

LUCIA NOCE
UNIVERSITY OF INSUBRIA,
DEPARTMENT OF THEORETICAL AND APPLIED SCIENCE,
VIA MAZZINI, 5, 21100 VARESE, ITALY.
E-MAIL: <LUCIA.NOCE@UNINSUBRIA.IT>

IGNAZIO GALLO
UNIVERSITY OF INSUBRIA,
DEPARTMENT OF THEORETICAL AND APPLIED SCIENCE,
VIA MAZZINI, 5, 21100 VARESE, ITALY.
E-MAIL: <IGNAZIO.GALLO@UNINSUBRIA.IT>

ALESSANDRO ZAMBERLETTI
UNIVERSITY OF INSUBRIA,
DEPARTMENT OF THEORETICAL AND APPLIED SCIENCE,
VIA MAZZINI, 5, 21100 VARESE, ITALY.
E-MAIL: <A.ZAMBERLETTI@UNINSUBRIA.IT>