2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Imbalanced Pattern Recognition: Concepts and Evaluations

Wladyslaw Homenda
Faculty of Mathematics and Information Science
Warsaw University Technology
ul. Koszykowa 75, 00–662 Warsaw, Poland
Faculty of Mathematics and Computer Science
University of Bialystok
ul. Sosnowa 64, 15-887 Bialystok, Poland
e–mail: homenda@mini.pw.edu.pl

Wojciech Lesinski
Faculty of Mathematics and Computer Science
University of Bialystok
ul. Sosnowa 64, 15-887 Bialystok, Poland
e–mail: wlesinski@ii.uwb.edu.pl

***Abstract*—In this paper we propose and investigate a concept of imbalanced pattern recognition problems and evaluation methods of solutions applied to solve such problems. The attention is focused on so called paper-to-computer technologies, but it is not limited to them due to possible direct generalization to other domains. Besides bringing a concept of imbalanced pattern recognition problem, classification quality from the perspective of single classes is considered. Parameters of binary classification and parameters and measures used in signal detection theory are adopted. Quality of classification in terms of one class contra all others is taken into account. Then, classifiers performance in frames of one class at the background of other classes and in frames of impact of other classes on the given on are evaluated. Finally, parameters characterizing global properties of classification are introduced and illustrated.**

I. Introduction

Pattern recognition problem is well elaborated for decades and from different viewpoints. The volume of researches is huge. It has implications in many areas. Its practical significance cannot be overestimated. Due to wide spectrum of applications in real life, there are still new aspects to be studied and resolved in both theory and practice. The aspect that is worth attention is imbalance of the problem. There are many real world data spaces, which are significantly asymmetrical, irregular, sporadic etc. Splitting population to healthy part and suffering from a given disease part is a standard example of imbalanced problem. The part of population suffering from a disease is much smaller than the healthy one. It is of great importance to identify ill individuals, so then we may admit to incorrect identification of healthy ones. On the other hand, a rate of incorrectly identified healthy individuals cannot be too big. In this study we consider imbalanced data to be recognized with importance given to multi classes pattern recognition problems. We formulate several quality parameters and then illustrate them with a real life problem of optical music recognition, an example of so called paper-to-computer technologies. Focusing attention on domain related problem does not limit considerations towards its direct generalization.

The paper is structured as follows. In Section II we recall background basic for the study. Concepts of imbalanced pattern recognition problem is presented in Section III. Evaluation of classification quality is given in section IV. Section V reflects the developed evaluation methodology in a problem of optical music recognition. Finally, conclusions close the discussion.

II. Preliminaries

Standard pattern recognition problem is a task of splitting a set of objects $\mathbb{O} = \{o_1, o_2, \ldots, o_{|\mathbb{O}|}\}$ into subsets, which include objects of the same class. Let us assume that the set of objects is split into m subsets, named classes, i.e. $\mathbb{O} = \cup_{i=1}^{m} O_i$ such that $(\forall i, j \in 1, 2, \ldots, m, i \neq j)\ O_i \cap O_j = \emptyset$. Such the task is defined by a mapping $\Psi : \mathbb{O} \to \mathbb{C}$ called *classifier*, where $\mathbb{O}$ is a set of objects and $\mathbb{C} = \{O_1, O_2, \ldots, O_m\}$ is a set of classes. For the sake of simplicity we assume that the mapping Ψ takes values from the set of indexes of classes $i \in M = \{1, \ldots m\}$ as its values instead of classes themselves.

Pattern recognition is usually performed on observed features, which characterize objects, rather then on objects directly. Therefore, we distinguish a mapping from the space of objects $\mathbb{O}$ into the space features $\mathbb{X}$, i.e. $\phi : \mathbb{O} \to \mathbb{X}$. This mapping is called *features extractor*. Then, we consider a mapping form the space of features into the space of classes $\psi : \mathbb{X} \to \mathbb{C}$. Such a mapping is named *classifier*. It is worth to notice that the term classifier is used in different contexts: classification of objects and classification of features. Meaning of this term can be concluded from context. Therefore, we will not distinguish explicitly, which meaning is taken.

Composition of the above two mappings constitute the classifier: $\Psi = \phi \circ \psi$. In other words, the mapping $\mathbb{O} \xrightarrow{\Psi} \mathbb{C}$ is decomposed to $\mathbb{O} \xrightarrow{\phi} \mathbb{X} \xrightarrow{\psi} \mathbb{C}$.

The space of features $\mathbb{X}$ is usually the Cartesian product of features $X_1, X_2, \ldots, X_n$, i.e. $\mathbb{X} = X_1 \times X_2 \times \ldots \times X_n$. Therefore, the mapping ϕ and ψ operate on vectors $x_1, x_2, \ldots, x_n$ (as values and arguments, respectively), where x_i is a value of the feature X_i for $i = 1, 2, \ldots, n$. For simplicity, a vector of values of features will be simply called a vector of features.

Usually, a classifier Ψ is not known, i.e. we do not know the class, to which an object belongs to. Finding such the classifier based on a learning set is the goal of a pattern recognition task. A learning set is a subset of the set of objects, $\mathbb{L} \subset \mathbb{O}$, for which classes are known, i.e. for any object from the learning set $o \in \mathbb{L}$ we know the value $\Psi(o)$.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

Summarizing, we explore pattern recognition problem searching for an (object) classifier:

$$\Psi : \mathbb{O} \to \mathbb{C}$$

assuming that it is known for $\mathbb{L} \subset \mathbb{O}$. Such the classifier is decomposed to a feature extractor:

$$\phi : \mathbb{O} \to \mathbb{X}$$

and a (features) classifier or a classification algorithm:

$$\psi : \mathbb{X} \to \mathbb{C}$$

Both the feature extractor and the classification algorithm are built based on the learning set $\mathbb{L}$.

The classifier ψ divides features' space onto so-called decision regions:

$$D_X^{(i)} = \psi^{-1}(i) = \{x \in X : \psi(x) = i\} \quad \text{for every} \quad i \in M$$

and then, of course, the features extractor splits the space of objects into classes:

$$O_i = \phi^{-1}\big(X^{(i)}\big) = \{o \in \mathbb{O} : \phi(o) \in X^{(i)}\} \quad \text{for} \quad i \in M$$

or equivalently:

$$O_i = \Psi^{-1}(i) = (\psi \circ \phi)^{-1}(i) = \phi^{-1}\big(\psi^{-1}(i)\big) \quad \text{for} \quad i \in M$$

We assume that classification algorithm splits the space of features' values, i.e. it separates the whole space X into pairwise disjoint subsets, which cover the whole space X:

$$(\forall i, j \in M, i \neq j) \;\; D_X^{(i)} \cap D_X^{(j)} = \emptyset \quad \text{and} \quad \bigcup_{i \in M} D_X^{(i)} = X$$

Pattern recognition problem not always has accurate symbols' extraction (segmentation) stage. Segmentation and extraction steps often produce many extraordinary undesirable symbols and ordinary garbage, let us call them *foreign symbols* in contrast to *native symbols* of recognized classes, c.f. [8]. In such a case a classification module, which assigns all extracted symbols to designed classes, will produce misclassification for every undesirable symbol and for every garbage symbol. Improvements of classification require construction of such classifiers which could assign designed symbols to correct classes and reject undesirable and garbage symbols.

Rejection of symbols can formally be interpreted in terms of a new class O_0, into which all undesirable and garbage symbols fall. Then we can distinguish a decision region, which separates foreign symbols from useful ones through the classifier ψ:

$$D_X^{(0)} = \{x \in \mathbb{X} : \psi(x) = 0\}$$

This new class (decision region) $D_X^{(0)}$ creates a new split of the space X

$$(\forall i \in M) \;\; D_X^{(i)} \cap D_X^{(0)} = \emptyset \quad \text{and} \quad \mathbb{X} = D_X^{(0)} \cup \bigcup_{i \in M} D_X^{(i)}$$

where, of course, all former classes $D_X^{(i)}, \;\; i \in M$ are pairwise disjoint.

Rejecting foreign symbols raises a problem since, unlike symbols of recognized classes, they are not similar and do not create a consistent class. Moreover, they are often not available at the stage of classifiers' designing. Therefore, instead of distinguishing a decision region corresponding to a class of foreign symbols, it is reasonable to separate areas outside of decision regions of native symbols, c.f. [8]. Of course, in such the case, we assume that decision regions of native symbols cover only their own areas and do not exhaust the whole feature space $\mathbb{X}$:

$$D_X^{(0)} = \mathbb{X} \setminus \bigcup_{i \in M} D_X^{(i)}$$

Rejection problem is an interesting topic, but we will not develop it in this work.

III. Imbalanced pattern recognition

A. Concepts and examples

1) Formulation: In this study we focus attention on paper documents' pattern recognition, which is a good example of paper-to-computer technology. However, discussed issues can be applied in different domains of pattern recognition applications. The following five characteristics of pattern recognition problem are regarded as imbalanced attributes:

- different cardinality of sets of symbols, some classes are heavily underrepresented, other are overrepresented,
- wide range of symbols' sizes,
- variability of symbols' shapes,
- irregularity of symbols' placement on the document,
- symbols of interest are overlapped by elements of the document and recognized symbols.

There are two standard problems, which can be used to explain practically imbalance: recognition of printed texts, known also as OCR problem, and recognition of printed music notation, known also as OMR.

2) OCR as balanced pattern recognition problem: According to the above description OCR problem is almost perfectly balanced, except the first attribute. Symbols of printed text are letters, digits, punctuation marks and some other symbols as, for example, algebraic operators, compare this text as example. The number of classes does not exceed 100. The following attributes are observed:

- cardinality of classes demonstrate wide range. Even if the set of classes is limited to symbols of a given document, e.g. a novel, we still have classes of different numbers of elements. For example, letters "a" and "e" create numerous classes while classes of "Q" and "Z" include small numbers of instances. From the perspective of this attribute, OCR problem is imbalanced,
- as in the previous point, sizes of symbols of printed texts do not show bigger difference. We can distinguish a few different sizes represented by lowercase and uppercase letters and punctuation marks. Anyway,

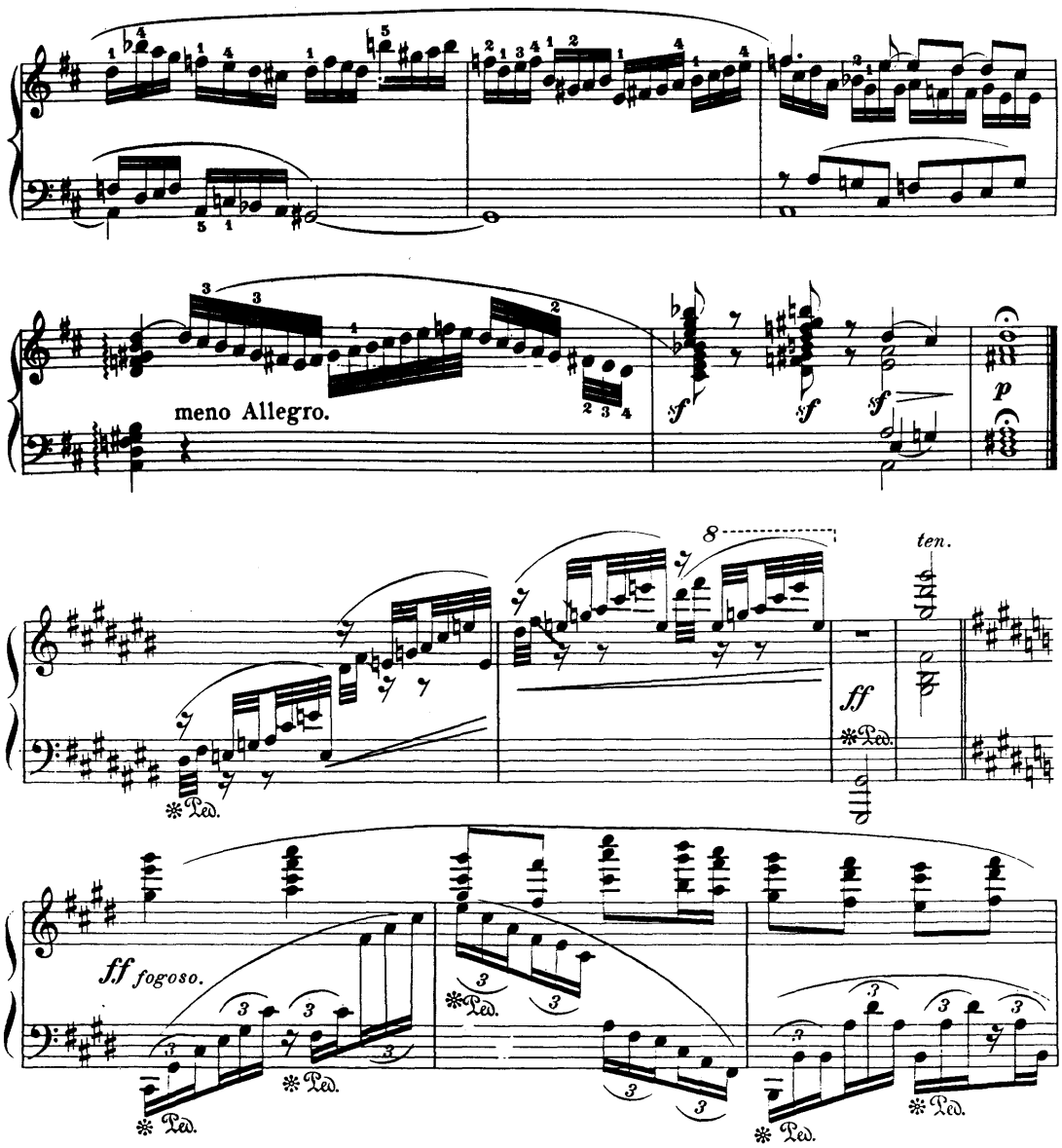

Fig. 1. Excerpts of typical piano scores with irregularity of symbols placement on the document.

proportion between sizes are regular and do not vary besides these few values,

- it is a matter of intuition what complexity of shapes is. Of course, shapes of letters ”a” and ”L” are different, but level of complication is not big. Therefore, based on commonsense, we can state that shapes of symbols do not show bigger variability,
- printed texts are examples of documents with perfectly regular placements of symbols, i.e. symbols are arranged in separated lines and are separated each from other in lines,
- no overlapping appears in printed texts. Sometimes, dirtiness, distortions of printing or careless scanning may result in gluing symbols.

3) OMR as imbalanced pattern recognition problem: OMR problem is an example of imbalanced pattern recognition

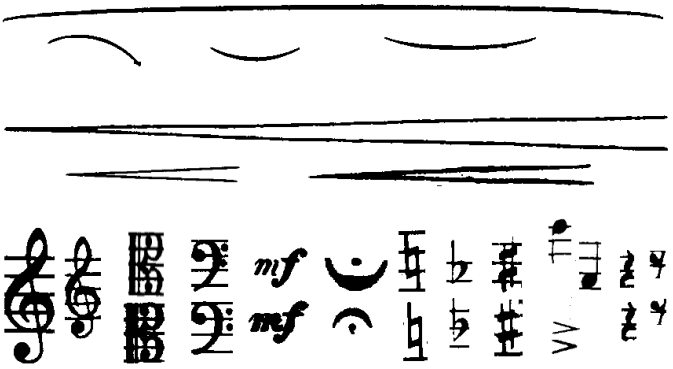

Fig. 2. OMR symbols: arcs (slurs and ties), dynamic markings (crescendo), clefs (G, C, F), mezzo forte, fermata, naturals, flats and sharps, quarter notes, accents, fourth and eight rests.

problem with regard to all five attributes. Both problems OCR and OMR are comparable with regard to number of classes, c.f. [6]. However, they differ in four out of five identified balanced attributes.

- cardinality of classes demonstrate wide range. We can call flats naturals or sharps as numerous classes. On the other hand, fermata, 32 rest, coda, dal segno symbols create small classes,
- symbols of printed music notation differs in size, c.f. top symbols and bottom left symbols of Figure 2,
- with regard to highly imbalanced shapes, it is sufficient to compare at Figure 2: top symbols (arcs and hairpins) and bottom right symbols (accents, notes) on one hand and bottom left symbols (clefs) on the other hand,
- symbols of printed scores are placed irregularly, what can be seen at Figure 1,
- symbols of printed scores are often overlapped or touched by other symbols or elements of music notation, which are not considered as goals of pattern recognition problem. Staff lines are the most noticable elements of music notation overlapping classified symbols, c.f. Figures 2 and 1.

B. Balancing

Such imbalance attributes as class cardinalities and symbol sizes can be subjected to some sort of balancing in order to turn an imbalanced problem to a balanced one. However, there is no consensus with regard to correctness of forcing balance of the problem.

1) Balancing cardinalities of classes: There are attempts for balancing cardinalities of classes. Minority classes can be subjected to so called over-sampling by either expanding sources of data sets and mining only minority symbols from extra sources. For instance, in order to balance cardinality of OCR classes, more printed documents may be used to dig for rare symbols. Also, some techniques might be applied to replicate existing minority symbols with, for instance, distorting replicated symbols.

On the other hand, overrepresented classes might be subjected to removal instances of such classes. Selection of symbols to be removed may be based on redundancy of symbols. Eliminated may be overrepresented symbols, which are falsely classified. Also, heavily noised symbols can be considered as not representative for a class.

Finally, as it is stated in [4], several inquiries state the convenience of apply- ing the under-sampling strategies when the level of imbalance is very low.

2) Balancing sizes of symbols: Objects of a pattern recognition problem, which are imbalanced with regard to their sizes, are often subjected to so called normalization. Usually, symbols of imbalanced size are resized to fit assumed square or rectangle. Typically, symbols are represented as raster rectangles, hence it is easy to resize them maintaining width/height ratio. There are several methods producing good quality resized raster rectangle. Anyway, such an operation does not preserve details of shrunk symbols or distorts details of enlarged symbols. Therefore, normalization of symbols of wide range variety of symbols' sizes may be more harmful then advantageous.

On the other hand, variability of sizes may be desirable to perform raw split of symbols. This is true assuming that inside classes size of symbols do not change, but they change between classes. Such a desirable feature is often not satisfied, for instance, mentioned above printed music notation has a feature of a natural anarchy of all imbalance attributes.

C. Segmentation and features selection

Standard pattern recognition problem based usually deals with balanced subjective data. Symbols of such problem are usually regularly placed on a recognized document. Therefore, segmentation is relatively easy. Symbols have similar size and similar level of shape's complexity. Symbols can be easily separated form others and from document's elements. Optical character recognition (OCR) is a typical example of such problem. On the other hand, OMR problem exhibits highly irregular placement on the document, as outlined in Figure 1. This feature raises troubles in automatic image segmentation process. As a result, segmentation outcome includes undesirable symbols and ordinary garbage, which are then subjected to classification stage. Quality of segmentation affects overall quality of the problem of pattern recognition. This topic is important in frames of pattern recognition field. However, it is not discussed in this study.

The mapping $\phi : \mathbb{O} \rightarrow \mathbb{X}$ represents the stage of feature selection. Pattern recognition problem rarely operates on recognized patterns directly. It might be the case in, for instance, OCR, when recognition is accomplished on raster bitmaps representing recognized symbols. Usually, classification is performed on features of recognized objects. Such features are acquired as observed or measured properties of objects. A task of features acquiring is in fact indispensable stage of any pattern recognition problem. This task was extensively studied and we do not refer to it and to related topics, because it is of less significance for the discussion.

IV. Evaluation of solution

Evaluating classification methods applied to imbalanced pattern recognition problem is the principal goal of this research. First of all, classification quality from the perspective of single classes is considered. We adopt parameters of binary classification evaluation and parameters and quality measures used in signal detection theory. Since these parameters are widely utilized, we do not refer to original sources, but of course do not claim to letting these factors on. In this point we recall employment of these factors in imbalanced two classes problem, c.f. [4]. Our goal here is to study possible evaluation of *multi classes problem*. We take into account quality of classification in terms of *one class contra all others*. Then, we evaluate classifiers performance in frames of one class at the background of other classes. Finally, we come to parameters characterizing global properties of classification.

A. Two classes problem

Evaluating a single factor cannot expose classification quality. This is true in general as well as in the two-classes problem. For instance, not only important is the proportion of the number of correctly recognized symbols of a class to the number of all symbols of this class. Obviously, the number of symbols falsely accounted to this class affects intuitive meaning of quality. Especially, when we consider a class of small number of elements, falsely classified symbols significantly decrease intuitive evaluation of quality. Therefore, we should look for formal evaluations compatible with intuition. Let us recall that in the case of imbalanced two-classes problem, the minority class is called positive one while majority class - negative one.

Such intuitive measures, as indicated above, provide a simple way of describing a classifier's performance on a given data set. However, they can be deceiving in certain situations and are highly sensitive to changes in data. For example, consider a problem where only 1% of the instances are positive. In such a situation, a simple strategy of labelling all new objects as members of other classes would give a predictive accuracy of 99%, but failing on all positive cases. In [4] the following confusion matrix was used in evaluating classification quality of a two classes problem, c.f. Table . Parameters exposed in this Table were then used in defining several factors, which outline classification quality.

TABLE I. CONFIUSION MATRIX FOR *two classes* PROBLEM

	Positive prediction	Negative prediction
Positive class	True Positives (TP)	False Negatives (FN)
Negative class	False Positives (FP)	True Negatives (TN)

B. Multi classes problem

For better of classification quality measuring, let us first consider the following parameters of *multi classes problem*. The parameters given in Table II are numbers of elements of a testing set which have the following meaning:

- TP - the number of elements of the considered class correctly classified to this class,
- FN - the number of elements of the considered class incorrectly classified to other classes,
- FP - the number of elements of other classes incorrectly classified to the considered class,
- TN - the number of elements of other classes correctly classified to other classes (no matter, if correctly, or not).

TABLE II. CONFIUSION MATRIX FOR *multi classes* PROBLEM

	Classification to the class	Classification to other classes
The class	True Positives (TP)	False Negatives (FN)
Other classes	False Positives (FP)	True Negatives (TN)

In this study we consider multi class problem. Hence, parameters of *two classes problem* are turned to *one class contra all others*, *one class in the background of others* and *all classes* characterization.

C. Local characterization

Now, we apply measures widely known and used in data transmission. We also comprise naming convention used there. Original sources are not referred to, but of course without claiming to letting these measures on, besides the last one listed below, i.e. besides $Separability$. These factors were already be used in *two classes* problem, for instance c.f. [4].

There are three pairs of complimentary factors. In order to increase quality of classification, Accuracy, Sensitivity and Precision should be maximized or, equivalently, Error, Miss Rate and False Discovery Rate should be minimized. Hence, there is no need to analyze all factors. It is sufficient to focus on one type of them.

The following in-class factors measure classification effectiveness inside a given class, i.e. taken are proportions of correctly recognized elements to all ones in this class. This is local factor strictly limited to the given class:

$$Sensitivity = \frac{TP}{TP+FN} \qquad (1)$$
$$Miss\ Rate = \frac{FN}{TP+FN} = 1 - Sensitivity$$

Classifiers' performance in an individual class at the background of all classes is measured by the following two factors. These factors measure effectiveness of correct acceptation of the class' elements together with correct rejection of elements not belonging to this class. No attention is given to classifiers' behavior in other classes:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \qquad (2)$$
$$Error = \frac{FP+FN}{TP+FN+FP+TN} = 1 - Accuracy$$

The next two factors evaluate influence of other classes at positive classification to the given class:

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$
$$False\ Discovery\ Rate = \frac{FP}{TP+FP} = 1 - Precision$$

D. Global characterization

In case of multi class classification problem there is no one universal quality parameter. Alike in local characterization, variety of configurations of multi class problems requires different viewpoints on classification's efficiency.

Expansion of sensitivity / Miss rate factor over whole classes can be done in several ways. The first one counts correct positive classification to all classes against all elements in all classes:

$$Sensitivity = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FN_i)} \quad (4)$$

$$Miss\ Rate = \frac{\sum_{i=1}^{m} FN_i}{\sum_{i=1}^{m} (TP_i + FN_i)} = 1 - Sensitivity$$

This parameter favours classes of bigger cardinalities while influence of small ones is imperceptible. Its application is justified when numerous classes are of similar cardinalities and detection of elements of small ones is less important.

The following factor equalizes numerous classes with small ones. In practice, it favours small classes. For instance, when there is more small classes than big ones, good classifier's performance on small classes increases average factor too much.

$$Sensitivity = \frac{1}{m} \sum_{i=1}^{m} \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$Miss\ Rate = \frac{1}{m} \sum_{i=1}^{m} \frac{FN_i}{TP_i + FN_i} = 1 - Sensitivity$$

The next measure is a case of the worst case (pessimistic) measure:

$$Sensitivity = \min \left\{ \frac{TP_i}{TP_i + FN_i} : i = 1, 2, \ldots, m \right\} \quad (6)$$

$$Miss\ Rate = \max \left\{ \frac{FN_i}{TP_i + FN_i} : i = 1, 2, \ldots, m \right\}$$

V. Illustrative example

This example is based on numerical results of an experiment performed for symbols of music notation. Synthesized are investigations described in [2], [7], [9] and [10].

A. Data set

In the experiment, the stage of image segmentation is not analyzed. The set of symbols was prepared as:

- partially as outcome of an automatic segmentation process of scanned scores. Then such symbols were subjected to necessary manual corrections and moved to appropriate classes,
- in other part, symbols were manually extracted,
- all symbols are represented as images of original sizes. They are standardized to a given size prior to classification.

Therefore, symbols subjected to design classifiers, i.e. belonging to training and testing sets could be seen as:

- being of similar size,
- having wide variation of shapes,
- suffering from being overlapped by other elements of document as, for example, staff lines.

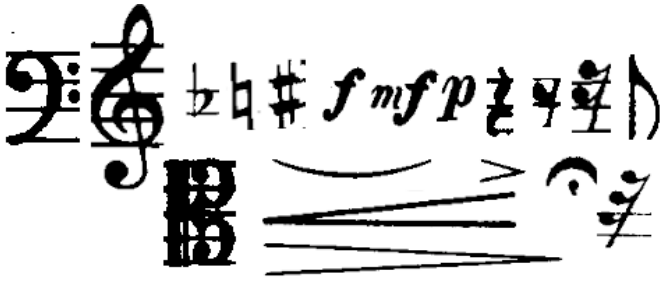

Fig. 3. Symbols being recognized: 1) numerous classes in upper row, left to right: clefs (F and G), chromatic symbols (flats, naturals and sharps), dynamic markings (forte, mezzo forte and piano), rests (quarter, eight, sixteenth), flagged stem, 2) rare symbols in bottom row, left to right, top to down: clef C, arc, crescendo, diminuendo, accent fermata, 32nd rest.

Two kind of classes are studied, c.f. Figure 3:

- numerous classes with 2000 elements each and
- small classes with different and relatively small number of elements: first four classes include 150-250 elements each and last three classes include 50-100 ones each, c.f. Figure 3.

B. Classification

Several classifiers were tested for classification symbols of printed scores. Tests were done in frames of [2], [3] and [10]. The following classifiers were utilized: simple classifiers (k-means clustering adopted for classification, classification based on Mahalanobis distance, decision tree and k-NN) and complex classifiers (voting, bagging and random forests). Aspects of classifiers' selection, single classifier tuning and behavior and details concerning comparative studies on selected classifiers are out of the scope of this paper, hence this topic is not developed here. For details see [10]. Only results illustrating evaluation of imbalanced recognition problem focused on recognition of printed music notation are considered.

C. Local characterization

Parameters of local characterization are outlined in Figure 4. The parameters concern the problem recalled here, i.e. recognition of printed music notation. Numerical results of symbols' classification are given in [10].

Accuracy measure for all recognized classes of symbols is shown in the top chart. Left twelve classes of symbols are numerous, right seven are classes of rare symbols. As mentioned before, Accuracy ia a measure counting correctly recognized symbols of a given class and correctly rejected ones from outside the class. No matter, if rejected symbols were accounted to correct classes or not. This measure favours numerous classes against small ones. This is why Accuracy gets almost 100% for small classes, c.f. right part of curves. Moreover, if number of balanced classes is relatively big, then share of a single numerous class in all classes is relatively small. Therefore, even a fair classifier gives high Accuracy

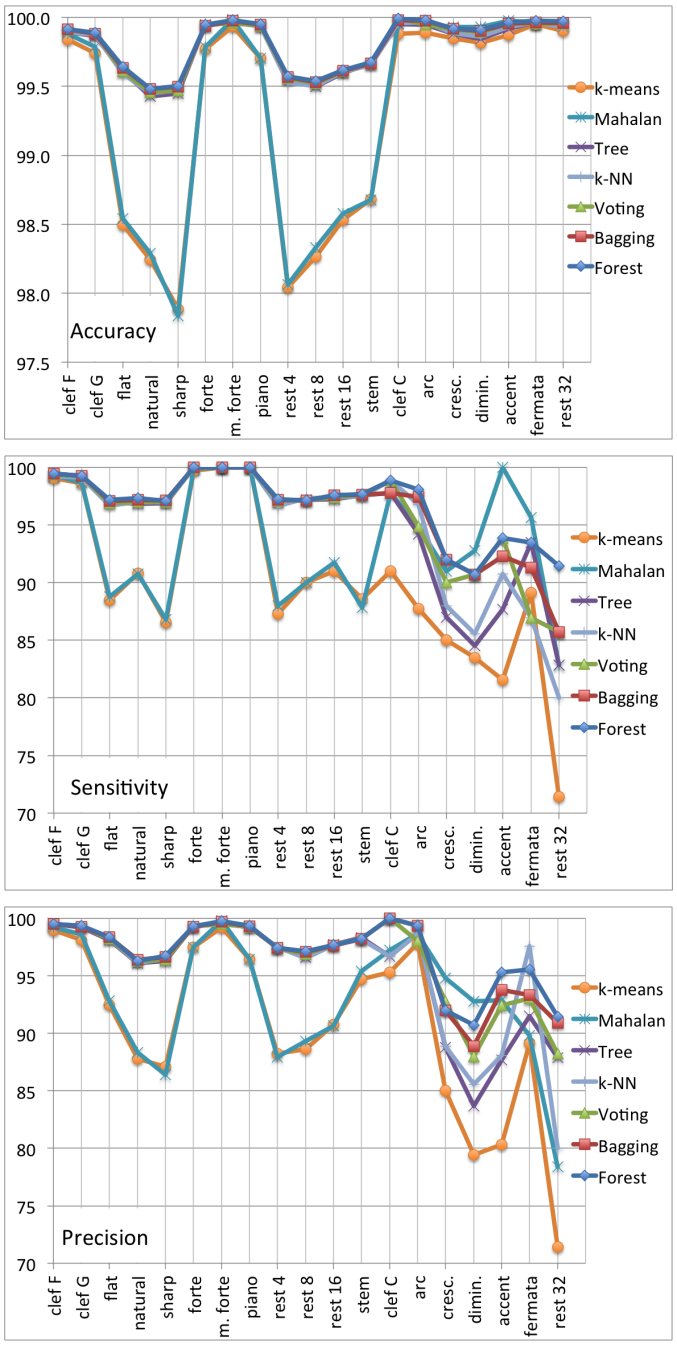

Fig. 4. Local characterization of classification for all classes of recognized symbols.

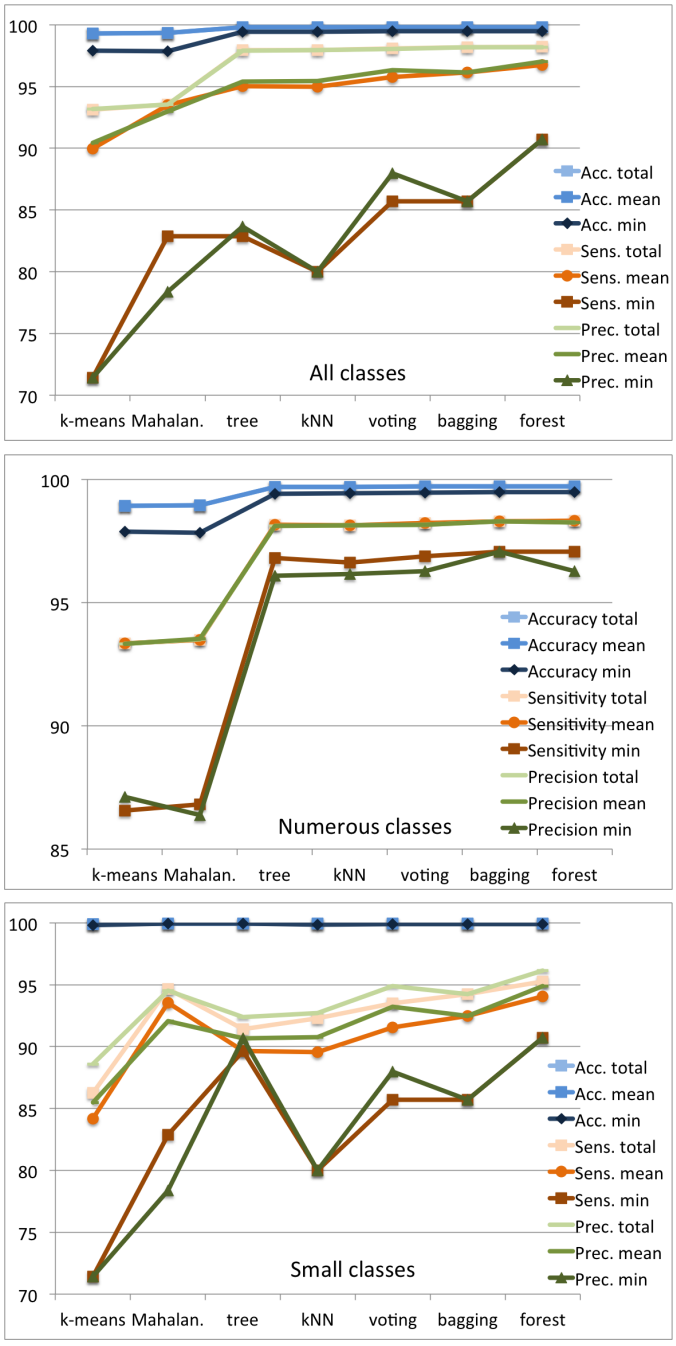

Fig. 5. Global parameters for all classes (top chart), numerous classes (middle chart) and small classes (bottom chart).

index. This feature can be observed in the left part of curves, where Accuracy index for most of tested classifiers gets 99.5% or more.

Accuracy index falls down for two classifiers (k-means and Mahalanobis) for two groups of symbols (chromatic symbols and rests and stem). The observed deterioration for all classifiers is raised by similarity of symbols inside these two groups. The deterioration is out of the scope of this study and is not discusses in details here.

Sensitivity and precision measures shown in Figure 4 have similar characteristic. firstly, for all classifiers indexes are worsen in small classes comparing to numerous ones. Anyway, both indexes are still high for the best classifiers exceeding 97% for numerous classes and 90% for small ones. As it is stated in Section IV-C, Sensitivity measures classifiers' performance inside single classes and Precision outlines impact of other class on a given one. Therefore, we can conclude that the best classifiers are doing well with correct classification of symbols for (all given) classes and with preventing false classifications to (all given) ones.

In the above discussion, we intentionally do not distinguish

classifiers referring, for instance, to the best ones. Analysis of given classifiers' behavior, direct or comparative, is out of interest of this discussion. Anyway, one may have a look at Figure 4 for such matters.

D. Global characterisation

Global parameters for evaluation imbalanced pattern recognition problem are defined in formulas 4, 5 and 6. Figure 5 illustrates evaluation of classification by measures formulated by these formulas. The upper chart of the Figure shows indexes for all classes, the middle chart illustrates indexes for numerous classes and the bottom one - for small classes.

Observation of charts in Figure 5 leads to some conclusions. Firstly, it is seen that Accuracy at global level exceeds two other measures (Sensitivity and Precision) and reaches 100% for small classes, c.f. blue curves. Secondly, the worst case characterization is the most demanding in all three groups of measures: Accuracy, Sensitivity and Precision. Curves for the worst case measures is placed below curves of two other measures. This observation is true for all three groups: Accuracy, Sensitivity and Precision. Thirdly, there are differences between performance of classifiers. Since analysis of concrete classifiers is out of the scope, it can only be stated that complex classifiers perform better than simple ones.

VI. Conclusion

In this paper we study imbalanced problems of pattern recognition with regard to their conceptualization and evaluations of solutions. We give an intuitive view on what imbalanced problem is. Some attributes of the statement of such problems are clear, e.t. cardinality of training/testing sets, variability of sizes or overlapping/intersecting of symbols and other elements of documents. On the other hand, variability of shapes of symbols is not defined and is left for intuition. A questions about shapes and and shapes' similarities/variabilities are worth a separate study from the perspective of pattern recognition. Alike, irregularity of symbols' placement on a document is not strictly defined as well. We illustrates all these attributes with examples, but not formally define them.

Then, we construct three measures for evaluation of different aspects of quality of classification in general and quality of classification of imbalanced pattern recognition problems. These measures are utilized at two levels: global and local. In order to be able to evaluate classification quality and to compare different classifiers, it is necessary to make analysis from different perspectives, since there is one universal measure for such evaluations.

Acknowledgment

The research is supported by the National Science Center, grant No 2012/07/B/ST6/01501, decision no UMO–2012/07/B/ST6/01501.

References

[1] Batista G.E., Prati R.C., and Monard M.C. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6:20 29, 2004

[2] Breaking accessibility barriers in information society. Braille Score - design and implementation of a computer program for processing music information for blind people, the research project no N R02 0019 06/2009 supported by by The National Center for Research and Development 2009-2012

[3] Cognitive maps with imperfect information as a tool of automatic data understanding. Ideas, methods, applications, the research project no 2011/01/B/ST6/06478 supported by the National Science Center, 2011-2014

[4] Garcia V., Sanchez J. S., Mollineda R. A., Alejo R., and Sotoca J. M. The class imbalance problem in pattern recognition and learning. In II Congreso Espanol de Informatica, pp. 283 291, 2007

[5] He, H. and Garcia, E. A., Learning from imbalanced data. Knowledge and Data Engineering, IEEE Transactions on, 21(9), 1263-1284, 2009.

[6] Homenda W. Optical Music Recognition: the Case Study of Pattern Recognition. Computer Recognition Systems. Springer Verlag, pp. 835-842, 2005.

[7] Homenda W., Lesinski W., Optical Music Recognition: Case of Pattern recognition with Undesirable and Garbage Symbols, in: Image Processing and Communications Challenges - Choras R. et al (Eds.), pp. 120-127, Exit, Warsaw, 2009

[8] Homenda W., Luckner M., Pedrycz W., Classification with rejection: concepts and formal evaluations, in: Andrzej M.J. Skulimowski (Ed.), Proceedings of KICSS'2013, pp. 161-172, Progress & Business Publishers, Krakow 2013

[9] Homenda W., Lesinski W., Features selection in character recognition with random forest classifier, Lecture Notes in Computer Science Volume 6922, 2011, pp 93-102

[10] Lesinski W., The Problem of Pattern Recognition for Imbalanced Subjects, PhD Dissertation, Warsaw, 2014.