KNN-Based Algorithm of Hard Case Detection in Datasets for Classification

Anton Okhrimenko[1] and Nataliia Kussul[1,2,3]

[1]*Institute of Physics and Technology, Igor Sikorsky Kyiv Polytechnic Institute, Peremohy Avenue 37, Kyiv, Ukraine*

[2]*Department of Space Information Technologies and System, Space Research Institute National Academy of Science of Ukraine an State Space Agency of Ukraine, Glushkov Avenue 40, Kyiv, Ukraine*

[3]*Anhalt University of Applied Sciences, Bernburger Str. 57, Köthen, Germany*

ant.okhrimenko@gmail.com, nataliia.kussul@gmail.com

Keywords: KNN, Dataset Quality Assessment, Imbalanced Datasets, Hard Cases.

Abstract: The machine learning models for classification are designed to find the best way to separate two or more classes. In case of class overlapping, there is no possible way to clearly separate such data. Any ML algorithm will fail to correctly classify a certain set of datapoints, which are surrounded by a significant number of another class data points at the feature space. However, being able to find such hardcases in a dataset allows using another set of rules than for normal data samples. In this work, we introduce a KNN-based detection algorithm of data points and subspaces for which the classification decision is ambiguous. The algorithm described in details along with demonstration on artificially generated dataset. Also, the possible usecases are discussed, including dataset quality assessment, custom ensemble strategy and data sampling modifications. The proposed algorithm can be used during full cycle of machine learning model developing, from forming train dataset to real case model inference.

1 INTRODUCTION

In various machine learning tasks a size and an quality of training data have a huge impact on its performance. For general case, the more data, the best model. But the data quality is important, we can't just duplicate our data or continuously sample it from the same source or object. Data must be diverse to cover the biggest possible volume at the feature space. This lead us two the second important requirement for training data: its quality. Speaking about classification task, the dataset must be separable, e.g. there must exist a surfaces which clearly separate data point and which are smooth enough to avoid overfitting. That means that data point forms some kind of clusters, each of which contains same-class data points only.

However, for the class overlapping problem the clear and unambiguous dividing surfaces can not be composited. Similar, the data point can not be distributed to the separate clusters. As result, a part of data point can not be clearly distinguished at the feature space and there are some subareas which contains a mix of different class points without any logic or structure. So typically the researcher faces the tasks to determine if given dataset fits for given classification task, to correct data acquisition process and, in case of classes overlapping problem still exists, to achieve the best possible result using given dataset.

For deeper problem understanding lets answer an question, why does the classes overlapping problem raise? It could be caused by an inaccurate data collection and/or labeling; or by improper data representation (lack of data complexity). In the last case, adding more dimensions to the feature space could greatly improve the data fitness for the given task. As a result, the researchers need an algorithm to determine weak points of the datasets, hard case percentage and subspaces with ambiguous data. Having such an algorithm, it becomes possible to make a decision to modify data gathering process, to append new feature or sample more numerous and accurate data points within and around questionable subspaces.

The most simple dataset investigation method is its visualization on two-dimensional plot. Large-scale datasets with numerous features can not be visualized directly without additional transformations. In this case, algorithms like PCA [1], tSNE [2] can be used for converting a large number of original features into a smaller set of converted features, and plotting it in

2D figure. Those methods have its weak points, such as human subjectivity and inability to effectively visualize and analyze high dimensional data due to high losses during the conversion.

However, in most cases researcher can not influence the dataset collection process and forced to work with partially overlapping data. In this case he also needs an algorithm, which will predict potential classification accuracy on given dataset and define unreliable data samples to correct model learning process.

At last, defined at the previous steps questionable subspaces at the feature space could be used at the inference time, making prediction using another set of rules, than for data which fall in reliable part of feature space.

So, hard case detection in feature space is very important scientific problem for image recognition quality. Reliable hard case extraction would improve quality of image recognition with convolutional neural networks (CNN). Typically CNN feature extractor converts image to embedding, which is a vector at the feature space. Then, the classification using this vector is performed. The problem there is that CNN is subject to change and class separation depends not only on data quality but on feature extractor quality.

It this study we propose a novel algorithm for hard case detection based on KNN classifier and discus its possible usages during machine learning model development.

2 RELATED WORKS

There are numerous researches conducted on dataset overlapping problem. Most of them are focused on the training stage enhancing providing methods to smooth the influence of dataset imperfectness. In general, class imbalance issue can be moderated by completely ignoring the data at the overlapping region, ignoring a majority class at the overlapping region or by making a separate rules for trusted and untrusted data [3]. This demonstrates us the importance of reliable method of overlaping area detection.

There are researches on KNN algorithm performance on overlapping and imbalanced datasets [4]. During the experiments generated datasets were used. The imbalance and overlapping percentages were main hyper-parameters for its generation as well as classes ratios at the overlapping zone.

Work [5] describes a method to deal with both class overlapping and imbalanced dataset problems using data undersampling. A result of two issues combinations is as follow: if class overlapping problem exists in imbalanced dataset, the machine learning model tends to classify all the data point in a controversial zones to the majority class, while ignoring the minority classes. Authors propose to determine data points of majority class which are lying near the data points on minority classes using KNN-based approach. Later those point are removed from the training dataset in order to balance minority and majority classes. This way model is trained to pay almost equal attention to all represented classes.

Authors of [6] also investigate the case of simultaneous class overlapping and class imbalance problem. In a similar way to the previous work they propose to delete majority class at the overlapping zone. The next stage is generating of the new datasets for several ensemble models by random balanced data sampling. The final model makes prediction based on a number of simple classifiers, among which each one has been learned on its own unique dataset.

Another way to deal with a class imbalance problem is to make a separate classification for different areas in the feature space. In [7] researchers propose to train two classification model, first one makes a binary decision about belonging to non-overlapping zones and second one based on SVM makes a classification decision in case of non-overlapping region. Several researches introduce two models, first one for overlapping area and second one for non overlapping.

Similar approach can be also used for dynamic model selection from an ensemble during the inference time [8]. Aside from class overlapping zones there are ”local unfair” zones where most models fail to make a correct prediction and as a result the final model fails also. However there are still models which make correct prediction. Thus, authors of [8] have proposed solution to detect such ones and select a subset of models that suit this zone. This way final model will have better metrics and comparable performance across the all feature space.

3 PROPOSED ALGORITHM OF HARD CASE DETECTION

Two possible problems that can arise during solving a classification task are classes overlapping and outliers (Figure 1). We propose algorithm which addresses with both problem simultaneously by detecting hard cases in dataset: data points, which any classification model will fail to recognize correctly with high probability. Finding such untrusted data points in the dataset will make great help for researcher during whole cycle of machine learning model developing.

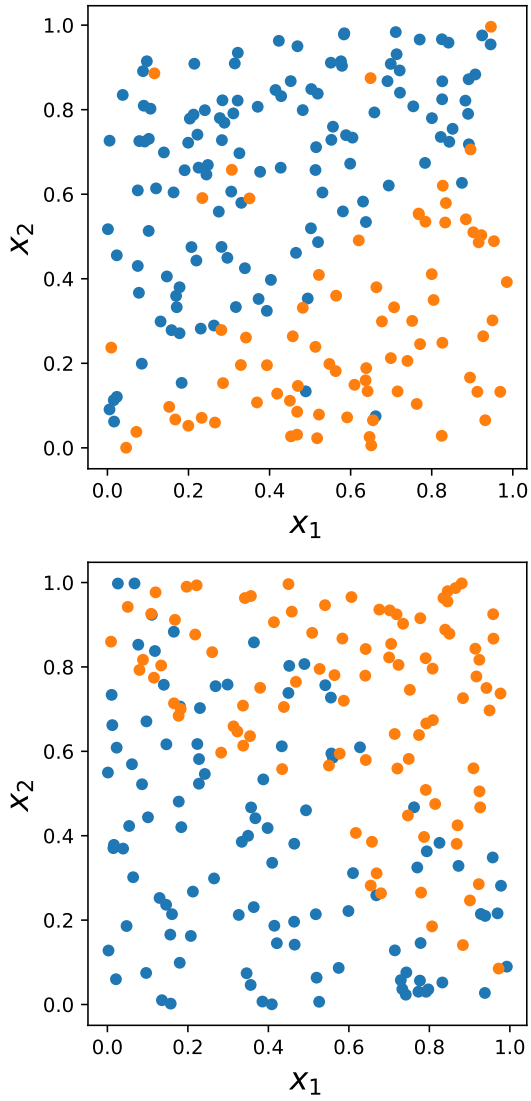

Figure 1: The outliers problem (up) and class overlapping problem (bottom).

3.1 Dataset Generation

In our study an artificially generated dataset is be used (Figure 2). It has two-dimensional features and three classes with a significant overlapping. Also several outliers are present. Three classes were sampled using random normal distribution $N(0.5, 0.2)$, $N(0.3, 0.3)$ and $N(0.2, 0.5)$. Values in range $[0.0, 2.5]$ are ignored.

For demonstration purpose we utilize two-dimensional dataset. Despite the small number of dimensions, the proposed algorithm works with any-dimensional data.

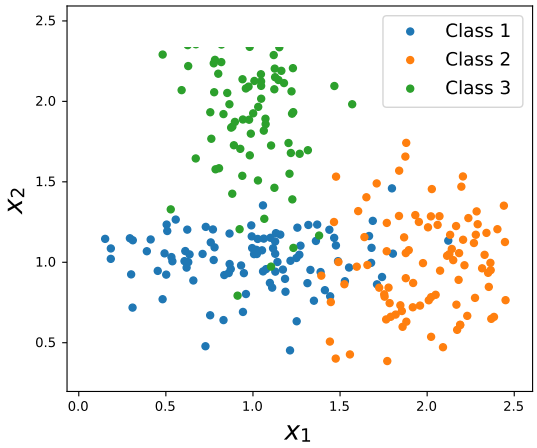

Figure 2: The generated dataset, that consists of three classes with class overlapping.

3.2 Hard Case Detection Algorithm

Let us define $\hat{X}$ as a set of data samples and $\vec{x}_i \in \hat{X}$ as a single data sample from this set. Similarly, lets define $\hat{Y}$ as a set of all possible classes and $y_i \in \hat{y}$ as a ground truth class for data sample $\vec{x}_i$.

We need to answer the question if it is possible for some data sample $\vec{x}_i$ to be correctly classified as class y_i. To do so we will use an ensemble of KNN classifiers with different neighbor number $n = [0, 1, \ldots, N], n \in \mathbb{N}$. For each $\vec{x}_i \in \hat{X}$ we get a vector $\vec{m}_i$ where m_i^j element is the result of classification of $\vec{m}_i$ by KNN classifier with neighbor number j which was trained using dataset $\hat{X} \setminus \vec{x}_i$.

$$\vec{m}_i : m_i^j = KNN(\vec{x}_i, j, \hat{X} \setminus \vec{x}_i) \tag{1}$$

This way each $\vec{x}_i$ will have its corresponding vector $\vec{m}_i$ and it is possible to construct a matrix M from this vectors:

$$M : M_{ij} = \{m_i^j\} \tag{2}$$

Now we can compare each vector $\vec{m}_i$ with groundtruth class y_i.

There are several cases:

- most elements of $\vec{m}_i$ match the true class y_i;
- first elements of $\vec{m}_i$ match the true class y_i, rest don't;
- most elements of $\vec{m}_i$ don't match the true class y_i;
- the predicted class m_i^j constantly changes depending on j, so-called class jumping;

To set a data sample as reliable we regard first two situations as obligatory. In fact the first one always

true when the second one is true, so there is one condition remains.

The data sample can't be reliable if third or fourth situation is present. The first case means that this sample is probably an outlier and the last one mean that data sample is surrounded with other sample with another class labels and probably within an overlapping zone.

Thus, the first condition C_1: from first k elements of vector $\vec{m}_i$ at least r must be equal to the right class y_i. k and r are hyperparameters and quite small. In general case, best values depends on dataset density.

The second condition C_2: data sample is unreliable if the most frequent class from first k elements of vector $\vec{m}_i$ isn't y_i.

The last condition C_3: if frequent class switching is present, data sample is unreliable. To put in algorithmic form lust check it with $1D$ convolution along the vector with kernel $K = [-1, 1]$ If two element are same, the convolution result will be zero. And for each case of class switching convolution result will be non-zero. Sum of the convolution result must not exec some value q which must be low.

The final rule for data sample classification is as follow:

$$C_1 \wedge \bar{C_2} \wedge \bar{C_3} \tag{3}$$

There are hyperparameters: k, r and q. Varying its values we can make some condition more important then another. As a general rule $q < r$ and $q < k$.

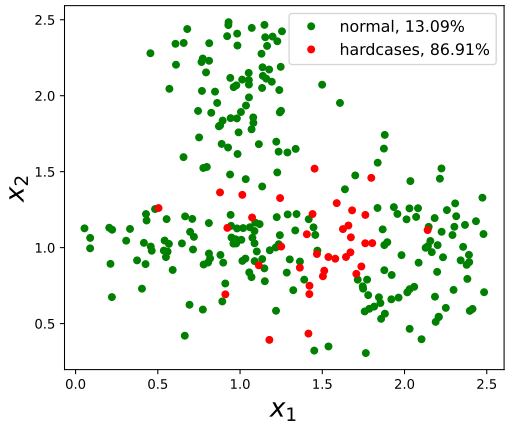

Figure 3: Detected hardcases, marked with a red color. Given a fraction of normal and hardcase data point in relation to a whole dataset size.

The example algorithm output is shown on Figure 3. It correctly detect most of questionable data points, although it has some problems with extreme data points.

4 PRACTICAL APPLICATION POSSIBILITIES

An algorithm of finding unreliable data samples is able to become one more tool at the researcher's toolbox and can be used during whole cycle of machine learning model development, from data collection to model deployment.

At the next subsections possible usecases f the proposed algorithm are discussed in details.

4.1 Dataset Quality Assessment

Introduced algorithm can prove its usefulness during both dataset collection and dataset research processes.

Most of modern machine learning problems require a huge datasets with high feature dimension and tremendous number of samples. Often there is no dataset that fits to current task. So many researchers are forced to create it, for example via outsourcing or crowd-sourcing. At this situation a tool to control the dataset quality is extremely important, because it can detect a problems at an early stages and give a possibility for a researcher to correct data gathering process.

Another situation take place when there are multiple datasets for a given problem. Usually the best ones are chosen or their combinations. Being able to determine a percentage of hardcases at given datasets, researcher can choose the dataset with the best quality. Also it possible to construct a new dataset from several ones with lowest hard case number aiming to minimize a fraction of unreliable data samples at united dataset.

4.2 Training Data Sampling Correction

As soon as it is possible to determine reliable and unreliable data sample at the training dataset, a lot of possible training enchantments are arisen. The first and the most oblivious one is to modify data sampling strategy.

During the model training, one can use reliable data for training input more frequently then unreliable ones. Balancing the frequency allow to feed the model is desired proportion, up to completely removing hard cases data. The resulting model will make decision more (heavily) based on reliable data samples rather than on unreliable ones. This allows it to make correct prediction with a high confidence at the non-overlapping zones and pay less attention to overlapping zones, because it is no way to correctly determine class at such areas anyway.

4.3 Custom Ensemble Strategy

Being able to clearly distinguish different data sample types by its quality gives an opportunity to train several models, each of which runs at its own areas of the feature space. This way the classification rules for zone with reliable data samples will be different from unreliable ones. Moreover, the models can be set to demand less confidence from the point at overlapping area during the training stage. The final model will consist of two subsets: first one for trusted data, and second one for data, which lies nearly the unreliable data points at the feature space.

However, the unreliable data sample can either lie at an overlapping zones, or to be an outlier. It is clear that the last case we can't threat with second set of classifiers and we need to filter such cases, so there is a room for improvements.

4.4 Dataset Modifications

Many methods of dealing with dataset imbalance and overlapping problems propose to delete majority class data sample at the overlapping zones [9]. To the contrary it is possible to make a class label modification without data deletion.

We need change class labels at the overlapping zones. To which class do we need to change labels is a subject for discussion and the answer is highly depends on current goals. For best possible metrics it make sense to change all class labels at the overlapping zone to the most frequent one. To fight an imbalance problem all class labels at questionable zone can be changed to the most underrepresented class.

As in the previous subsection, we need to filter outliers, which class labels can be changed to the labels of surrounding data points.

5 CONCLUSIONS AND FURTHER WORK

This work is dedicated to investigation of a class overlapping problem along with other possible hard cases such as outliers. There are many reasons for such a problem and the most important are inaccurate data sampling and lack of data dimension (e.g. there is a need to increase the dimension of feature space). Regardless of the reasons, there are some points which simply can't be classified correctly, because they have similar features with another data samples with different class label. As an example we can consider thematic segmentation problem to determine land use based on satellite data [10]. Due to similar spectral characteristics of such crops as wheat and barley, it is impossible to separate them at the crop specific map. Such data points interfere the model training process and could become a source of incorrect prediction on real data.

At this study a novel algorithm has been introduced, which allows any researcher to have more clear understanding of datasets quality and it does not depend on feature space dimension. Apart from visual estimation methods based on dataset decomposition, this method is able to produce a clear numeric data quality metrics such as the percentage of trustful data. More over, it allows to distinguish reliable data samples and unreliable ones, opening an opportunity to adjusting.

The proposed algorithm could make an enchantment during the whole cycle of machine learning model development. This include correction and adjustment of the dataset collection process, as well as dataset choosing and mixing. Having hard case detected, one can apply different rules for reliable and unreliable data during training and inference runs. The algorithm also opens room for dataset modification, according to which class labels of untrustworthy data samples are changed to match the current goals, such as high metrics or class balance.

In this study we consider artificially generated datasets, but our further steps will be related to utilization of this algorithm for real world problem of land use classification on satellite data.

REFERENCES

[1] H. Abdi and L. J. Williams, "Principal component analysis. wiley interdisciplinary reviews: computational statistics," Wiley Interdisplinary Reviews: Computational Statistics, 2010.

[2] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579-2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[3] W. A. Almutairi and R. Janicki, "On relationships between imbalance and overlapping of datasets," EPiC Series in Computing, vol. 69, 2020.

[4] V. García, R. A. Mollineda, and J. S. Sánchez, "On the k-nn performance in a challenging scenario of imbalance and overlapping," Pattern Analysis and Applications, vol. 11, 2008.

[5] M. M. Nwe and K. T. Lynn, "Knn-based overlapping samples filter approach for classification of imbalanced data," Studies in Computational Intelligence, vol. 845, 2020.

[6] L. Chen, B. Fang, Z. Shang, and Y. Tang, "Tackling class overlap and imbalance problems in software

defect prediction," Software Quality Journal, vol. 26, no. 1, pp. 97–125, Mar 2018. [Online]. Available: https://doi.org/10.1007/s11219-016-9342-6.

[7] Y. Tang and J. Gao, "Improved classification for problem involving overlapping patterns," IEICE TRANSACTIONS on Information and Systems", vol. 90, no. 11, pp. 1787–1795, Nov 2007. [Online]. Available: https://doi.org/10.1093/ietisy/e90-d.11.1787.

[8] N. Lässig, S. Oppold, and M. Herschel, "Metrics and algorithms for locally fair and accurate classifications using ensembles," Datenbank-Spektrum, vol. 22, 2022.

[9] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Comput. Surv., vol. 52, no. 4, aug 2019. [Online]. Available: https://doi.org/10.1145/3343440.

[10] N. Kussul, A. Shelestov, M. Lavreniuk, I. Butko, and S. Skakun, "Deep learning approach for large scale land cover mapping based on remote sensing data fusion," International Geoscience and Remote Sensing Symposium (IGARSS), vol. 2016-November, 2016.