The sampling method preserving interclass boundaries

Dmytro Kavrin[1][0000-0002-8952-4067], Sergey Subbotin[2][0000-0001-5814-8268],

[1,2]Zaporizhzhia National Technical University, Zhukovsky str., 64,
Zaporizhzhia, 69063, Ukraine
[1]kavrin@gmail.com, [2]subbotin@zntu.edu.ua

Abstract. The sample dimensionality reduction problem for classification is addressed. The sampling method with the preservation of the most significant instances near the interclass boundaries is proposed. It calculates the interclass distances in the sample used to build hyperspheres, removes redundant instances inside the hyperspheres, and creates a subsample from the set of hypersphere centers. The experiments to study the proposed method properties are conducted. They allow recommending the proposed method for use in practice as a significant to reduce the complexity and ensure acceptable accuracy of classification models.

Keywords: class, classification, interclass boundaries, cluster, diagnosis, instance, metric, sample.

1 Introduction

The decision making automation in the problems of technical and biomedical diagnostics requires the construction of classification model that usually created on a set of observations (instances, precedents) characterized by a set of features [1]. In applied problems often we need to operate with a large-size datasets. This leads to a significant expense of time for data processing and requires big computer memory resources. Therefore, the urgent problem is to reduce the data sample dimensionality.

There are two main approaches to solve a data dimensionality reduction problem: a feature selection and a sample selection [2].

The approach of feature selection is the most widely used in practice [3]. It enumerates various combinations of features and determines their suitability for a model building based on a specific criterion for feature combination. Among all the best feature combinations the resulting combination the one is selected as a result that contains the smallest feature number. However, if the feature set is not initially redundant, this approach may lead to the loss of important information.

Another approach is a sample (instance) selection using different methods to reduce the number of instances [4–7]. The most common practice is a random extraction of smaller subsamples from the big original sample. But random methods in most cases cannot guarantee that the resulting sample of a small size will reflect the main properties of the original sample, especially near the interclass boundaries.

The purpose of this paper is to develop a sample selection method, which allows minimizing the size of original samples and at the same time to preserve the most important instances located near the interclass boundaries.

2 Formal problem statement

Let we have an original data sample $X = \langle x, y \rangle$, where $x = \{x^s\}, y = \{y^s\}, s = 1,2,...,S$, where *S* is a number of instances, y^s is a class of *s*-th instance, $x^s = \{x_j^s\}$ is a *s*-th instance inputs, *j=1, 2, ..., N*, x_j^s is a value of *j*-th input feature x_j for *s*-th instance, *N* is a number of input features.

Then the problem of sample selection may be presented in the form: find $X' = \langle x', y' \rangle$: $x' \in \{x^s\}, y' = \{y^s \mid x^s \in x'\}, S' \leq S, f(\langle x', y' \rangle, \langle x, y \rangle) \rightarrow opt$, where *<x',y'>* is a selected subsample, *S'* is a number of instances in a selected subsample, *f()* is a criterion describing quality of sample selection, opt is an acceptable value of *f* criterion. As a rule, for the problems of approximation the model quality criterion is determined as a function of the model error.

3 Literature review

The sample dimensionality reduction methods by extracting smaller subsamples from the original samples can be divided into two main categories: probabilistic and deterministic methods [4–20].

Probabilistic methods involve randomly selecting of each instance (group of instances) from the original sample with a known non-zero probability that can be accurately determined. The probabilistic sampling [4–18] includes the following methods:

- simple random sampling method [10], which allows randomly selecting a given number of instances from the original sample. Moreover, all instances of the original sample have the same probability of being selected;
- systematic selection method [11], which arranges the original sample in a certain form and splits it into consecutive groups of instances. Then an object with a given sequence number from each group is selected and included in the subsample being formed;
- stratified selection method [12], which divides the original sample into non-intersecting homogeneous subsets (strata), including instances of all types. Then random or systematic selection methods are applied to each subset;
- probability proportional to size sampling method [13–14], which is applied when there is additional information about the classes and their size, and the probability of selecting each instance of the original sample will be proportional to the size of the class to which it belongs;

- cluster sampling method [15–18], which divides instances of the original sample into clusters. Then, from each cluster, a subset of the instances for the subsample is randomly selected.

The advantages of probabilistic methods [5, 10] are their relative simplicity and the possibility of estimating the sampling error. The disadvantages of probabilistic methods are that they do not guarantee that a subsample will display the properties of the original sample well or will not be redundant and will not artificially simplify the task.

Deterministic sampling methods [5, 19] involve the selection of instances based on assumptions about their informativeness, which forms the selection criteria. In this case, the sample data contains instances that may not be selected or the probability of their selection cannot be accurately determined. Therefore, the theory developed for probabilistic samples is not applicable to such samples. The deterministic sampling includes the following methods:

- convenience sampling [5, 20], which forms a non-representative sample of the instances most easily available for study;
- quota sampling [5], which divides the original sample into disjoint subgroups with different properties, after which instances are selected from each subgroup based on a given proportion and on the researcher's preferences;
- purposive sampling [20], which extracts instances from the original sample in accordance with the researcher's opinion about their relevance to the study.

A major problem with these methods [19–20] is the impossibility of estimating the error of the formed samples. The advantage of deterministic methods is their ability to identify the most significant instances for building a diagnostic model of precedents, which can also be used to initialize recognition patterns and speed up the learning process.

The original data samples used in solving the problem of building a diagnostic model by precedents can be very large or have redundant data. Model operation using such samples may require significant computational and time resources. Extracting smaller subsamples from the original data is an effective and natural solution to the big data problem when building a diagnostic model by precedents. It is important to preserve significant instances of the original sample to obtain a representative subsample of a smaller volume. Therefore, when constructing diagnostic models, based on cluster analysis deterministic methods are the most relevant, since they make it possible to identify the most significant instances. As a rule, when solving problems of technical and biomedical diagnosis there is information about classes, this greatly simplifies the task of clustering, which can be performed more accurately, especially near the class boundaries. However, in the common case, cluster analysis solves the clustering problem for unlabeled data and does not allow selecting important instances located near interclass boundaries.

Therefore, to create representative subsample of a small size with well-defined class boundaries, it is necessary to develop a method of sampling from the original data, which will reduce the computational load and preserve the topological representativeness of the original sample in the feature space by keeping significant instances at the interclass boundaries.

4 The sampling method preserving interclass boundaries

To preserve the topological representativeness of the original sample in the feature space by keeping significant instances at the interclass boundaries a deterministic method of labeled datasets reduction is proposed. It separates the original labeled sample by the hyperspheres of different radii in the feature space, depending on the distances between instances of different classes. The reduction is performed sequentially for each class (primary class) relative to all other classes. On each method's iteration, the most perspective instance of primary class is selected. This is calculated using the Gromov–Hausdorff distance between the primary class and the united set of other classes:

$$D_h = \arg \min_{p=1,2,\ldots,S_p} \left\{ d(x^p, x^u) \right\}, \tag{1}$$

where x^p is an instance of the primary class and the most perspective instance, x^u is an instance of the united set of other classes.

Then a hypersphere is formed with a center at the most perspective point and a radius equal to the distance to the nearest instance of the joint set of other classes in the feature space. All instances of primary class inside the hypersphere are excluding from further consideration. The procedure is repeated until all instances of the primary class will be excluded from the consideration. The reduced subsample is formed from the centers of the resulting hyperspheres. The construction of hyperspheres is performed separately for each class of the original sample. Thus, the method adapts to the distribution of data in the sample, automatically adjusting the number of instances in the reduced subsample. Due to the adaptability of the hypersphere radii, closer to the class boundary more hyperspheres of a smaller radius are formed. In other words, in a reduced sample, the density of instances at the class boundary will be higher than far from the boundaries, which allows for more accurate determination of the class boundaries.

The proposed method is based on the hypothesis of compactness of classes. In the ideal case, when classes do not intersect (compact), the proposed method greatly reduces the original sample, including by removing meaningful instances at interclass boundaries, thereby skewing interclass boundaries. In this case, in order to form a more representative subsample, it is proposed to regulate the number of hyperspheres by the fractional change in their radii. With a decrease in the hypersphere radii, the number of clusters increases, and the most of the clusters accumulate at the class boundaries. Thus, in the formed subsample, it is possible to regulate the number of instances, especially at the class boundary, increasing its representativeness. Formally, this method can be written as follows:

1. The initialization stage. Set the initial data sample $X = \langle x, y \rangle$ and initialize the reduced sample $X' = \langle x', y' \rangle = \varnothing$.
2. The stage of sample splitting into classes. Split the original sample $X = \langle x, y \rangle$ into K separate subsamples $X(k)$ for the instances of each class: $X(k) = \bigcup_{s=1}^{S} \{x^s \mid y^s = k\}$, where $k = 1,...,K$. Determine the volume (number of instances) of each k-th subsample S_k.
3. The stage of subsample reduction. Set $k = 0$. Then while $k < K$ perform in a cycle: $k = k + 1$, set primary subsample $P = X(k)$, merge all other subsamples into one virtual subsample $U = \bigcup_{s=1}^{S} \{x^s \mid y^s \neq k\}$, further while $P \neq \varnothing$ perform in a cycle:

- calculate the Gromov–Hausdorff distance between the primary subsample P and the virtual subsamples U:

$$d(x^p, x^u) = \arg \min_{p=1,2,...,S_p} \left\{ \sqrt{\sum_{j=1}^{N} (x_j^p - x_j^u)^2} \right\}; \qquad (2)$$

- set the instance x^p as the most perspective instance of the primary subsample P and join it to the subsample X': $X' = X' \bigcup x^p$;
- determine the radius of the hypersphere centered at x^p as $r = \lambda d(x^p, x^u)$, where λ is a fraction of distance d, $0 \leq \lambda \leq 1$;
- join the most perspective instance x^p to the subsample X': $X' = X' \bigcup x^p$;
- remove from the subsample P all instances that are included in the hypersphere of radius r: $P = P \setminus P^r$, where P^r is a set of instances remote from the center x^p of the hypersphere at a distance $r < d(x^a, x^s)$.

4. The stage of model building. Construct the recognition model using reduced sample X'.

The proposed method allows automating the process of reducing the size of the original sample, which contains information about the classes, for solving the problems of technical and biomedical diagnosis.

The disadvantage of the method is the necessity of calculating and memory storing of the pairwise distances between instances of opposite classes. Therefore, if the size of the original sample is large enough and does not allow all pairwise distances to be simultaneously loaded into the computer's memory, or the data are received dynamically, it is possible to process the original sample in packets.

To provide the correct work of the method each packet should be represented by all classes. Applying the described method to the first data packet, we obtain the initial distribution of the extracted sample. The instances of the next data packet are combined with the instances of the sample obtained in the previous step. The combined sample is processed again by the sampling method with preserves interclass boundaries. The procedure is repeated until all the data of the original sample has been processed, or until a stop has been made in the specified phase of the data stream, if the data is dynamic. Formally, this method can be written as follows:

1. The initialization stage. Set the initial data sample $X = \langle x, y \rangle$ and initialize the reduced sample $X' = \langle x', y' \rangle = \varnothing$.
2. The stage of packets initialization. Determine the number of packets in the sample $P = \text{round}(S / Q)$, where Q is a number of instances in the packet, specified by the user, round is an argument rounding function to the nearest integer. Split the original sample X into P packets $X(\rho)$, $\rho = 1,...,P$.
3. The stage of packet processing. Set $\rho = 0$. While $\rho < P$, perform in a cycle: set $\rho = \rho + 1$, join packet $X(\rho)$ with reducing sample X': $X(\rho) = X' \cup X(\rho)$, then process the packet $X(\rho)$ using the sampling method with preserving class boundaries described above using the adaptive reduction method and set $X' = X(\rho)$.
4. The model building stage. Construct the recognition model from a reduced sample X'.

The proposed batch sampling method allows to process of very large samples in a batch of a given size. The data batch processing combined with the sampling method preserving interclass boundaries allows to obtain representative reduced samples from large data samples, or dynamic data sets, without requiring significant computational resources.

5 Experiments

To study the properties of the proposed methods they were implemented as computer software. The set of experiments were conducted to solve the practical problems of classification using the developed software implementing proposed methods.

Considering the possible variability of relative performances of method across datasets, the results were obtained based on two datasets of various size and dimension, characterized in Table 1.

Table 1. Datasets used for validate the method

Dataset	Number of instances	Number of features	Number of classes
Breast Cancer Wisconsin (Diagnostic)	569	30	2
Activity recognition with healthy older people using a batteryless wearable sensor	9101	8	5

The Breast Cancer Wisconsin Diagnostic dataset (BSW) requires to predict whether the cancer is benign or malignant. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image [21].

The activity recognition with healthy older people using a batteryless wearable sensor (RHOP) dataset contains sequential motion data from 14 healthy older people aged 66 to 86 years old using a batteryless, wearable sensor on top of their clothing for the recognition of activities in clinical environments [22].

In each problem the data were normalized to map feature values into the interval [0, 1]:

$$x_j^s = \frac{x_j^s - x_j^{\min}}{x_j^{\max} - x_j^{\min}}, \tag{3}$$

where x_j^s is a j-th feature value of s-th instance of a sample, $x_j^{\min}$ is a minimum value of j-th feature, $x_j^{\max}$ is a maximum value of j-th feature.

Further, each initial data set was divided into test and training samples by the stratification method [12] in the ratio of 25% and 75%, respectively. This made it possible to compare the work of models built on the original sample and reduced sample by the proposed method.

The classification model was built using a two-layer feed-forward neural network with 10 neurons in a hidden layer trained by the Levenberg-Marquardt method and Error backpropogation technique [23].

For the comparison of models built on different training samples the mean square error of the test sample classification was calculated:

$$MSE = \frac{\sum_{s=1}^{S} \left(y^s - f(x^s)\right)^2}{S}, \tag{4}$$

To provide a reliable result in our study ten simulations of model building for each training sample were performed. After model testing their *MSE* values were averaged.

The results of the proposed method depend on the density of the instance distribution near the interclass boundaries and the class compactness. So, for a sample with heavily mixed classes, the size of the reduced subsample may remain almost unchanged, or it may change slightly. If the classes are well separable then the size of the reduced subsample will be small enough and the interclass boundaries will be poorly defined.

Therefore, the key issue of applying the proposed method is the correct selection of the λ coefficient. Selection of λ coefficient allows to change the radii of hyperspheres, thus adjusting the number of instances near the interclass boundaries and the representativeness of the reduced sample. In the experiments, a training set was created for different values of λ coefficient (0.25, 0.5, 0.75, 1). Then, the *MSE* dependences on the λ coefficient were studied.

6 Results and Discussion

The conducted experiments showed that the proposed method works well in automatic mode due to its adaptability, which is an important factor in solving technical and biomedical diagnostics problems.

The results of the proposed method are presented in Table 2.

Table 2. Results of experiments

λ	BSW		RHOP	
	S'	*MSE*	*S'*	*MSE*
0	427	0.026194	6902	0.003484
0.25	427	0.027152	623	0.010707
0.5	372	0.027807	329	0.007960
0.75	228	0.027599	201	0.012432
1	127	0.035558	136	0.026053

As it can be seen from the Table 2, the results for the different data sets are also different. This is due to the different distribution of instances at interclass boundaries.

In the case of BSW, the original size of the training sample was 427 instances. From Table 2 it can be seen that *MSE* has changed slightly for all λ values, while the sample size has reduced almost 4 times with $\lambda = 1$. Thus, using the proposed method for the BSW dataset, it was possible to significantly reduce the size and save the most important instances on the interclass boundaries.

The original size of the training sample of the RHOP dataset was 6902 instances. It can be seen that for given λ values, the sample size has significantly reduced, and the value of *MSE* has increased. This may mean that the classes are very well separated and the density of the instances at the interclass boundaries is low, so the use of the developed method led to a deficiency of instances at the interclass boundaries, respectively to a decrease in the model's efficiency. Therefore, in this case, it is possible to recommend to set the minimum λ values ($0 < \lambda \le 0.1$).

The proposed method makes possible to find a compromise between the size and representativeness of the reduced samples, depending on the tasks by changing the proportion of the hypersphere radii using the λ coefficient.

The disadvantage of the developed method is that it is computationally expensive, especially for large datasets.

Therefore, there is a need to apply approaches to reduce the method computational load. For example, in the case of large samples, it is possible to use the method in an ensemble with feature selection methods [24].

It is also possible to reduce the computational load by eliminating the stage of determining the most promising instance from the calculations and it's replace by a random instance of the class. However, this approach requires further study and identification of the necessary restrictions. The next approach could be parallelization of the computational load using multiprocessor systems.

7 Conclusion

The problem of reducing the labeled large data samples for diagnostic model building by precedents is addressed in the paper.

The scientific novelty of study results consists in the fact that a new sampling method preserving interclass boundaries has been developed. It makes possible to significantly reduce the size of the original labeled sample, retaining the most significant instances near the interclass boundaries and removing less informative instances located inside the classes. Thus, the proposed method allows in the automatic mode to solve the sample reduction problem adapting to the data distribution in the labeled sample.

The practical significance of the obtained results is that proposed method is implemented as software, which provides possibility of batch processing of large data samples, or samples formed from data streams (dynamically incoming data). This software has been experimentally studied at solving the problems of real datasets sample selection. The experiments were confirmed the efficiency of the developed software and of implemented method. The results of the conducted experiments allow to recommend the use of the developed method and its software implementation for solving the problems of technical and biomedical diagnosis.

The prospects for further research may be concerned on study the proposed method on a wider class of practical problems. Also the study of the possibilities of the proposed method in ensembles with the methods of feature selection for large datasets seems to be appropriate. The development of implementations of the proposed method for multiprocessor systems operating in a parallel mode also is important for many practical problems.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2016)
2. García, S., Luengo, J., Herrera, F.: Data Preprocessing in Data Mining. Springer, Switzerland (2016)
3. Khalid, S., Khalil, T., Nasreen, S.: A survey of feature selection and feature extraction techniques in machine learning. In: 2014 Science and Information Conference, pp. 372-378. IEEE Press, London (2014). doi: 10.1109/SAI.2014.6918213
4. Thompson, S.K.: Sampling. John Wiley & Sons, Hoboken (2012)
5. Lavrakas, P.J.: Encyclopedia of survey research methods. Sage Publications, Thousand Oaks (2008)
6. Cochran, W.G.: Sampling Techniques. John Wiley & Sons, New York (1977)
7. Chaudhuri, A., Stenger, H.: Survey sampling theory and method. Chapman & Hall, New York (2005)
8. Subbotin, S.A.: The sample properties evaluation for pattern recognition and intelligent diagnosis. In: The 10th International Conference on Digital Technologies 2014, pp. 321-332. IEEE Press, Zilina (2014). doi: 10.1109/DT.2014.6868734
9. Subbotin, S.A.: The training set quality measures for neural network learning. In: Optical Memory and Neural Networks (Information Optics), vol. 19(2), pp. 126-139. Allerton Press (2010). doi: 10.3103/S1060992X10020037

10. Tille, Y., Wilhelm, M.: Probability Sampling Designs: Principles for Choice of Design and Balancing. In: Statistical Science, vol. 32(2), pp. 176–189. Project Euclid (2017). doi:10.1214/16-STS606
11. Kalton, G.: Systematic Sampling. In: Wiley StatsRef: Statistics Reference Online. John Wiley & Sons (2017). doi: 10.1002/9781118445112.stat03380.pub2
12. Parsons, V.L.: Stratified Sampling. In: Wiley StatsRef: Statistics Reference Online. John Wiley & Sons (2017). doi: 10.1002/9781118445112.stat05999.pub2
13. Skinner, C.J.: Probability Proportional to Size (PPS) Sampling. In: Wiley StatsRef: Statistics Reference Online. John Wiley & Sons (2016). doi: 10.1002/9781118445112.stat03346.pub2
14. Subbotin, S.A.: Methods of sampling based on exhaustive and evolutionary search. In: Automatic Control and Computer Sciences, vol. 47(3), pp. 113-121. Allerton Press (2013). doi: 10.3103/S0146411613030073
15. Nelson, G.F.: Cluster Sampling: A Pervasive, Yet Little Recognized Survey Design in Fisheries Research. In: Transactions of the American Fisheries Society, vol. 143(4), pp. 926–938. John Wiley & Sons (2014). doi: 10.1080/00028487.2014.901252
16. Ly, T., Cockburn, M., Langholz, B.: Cost-efficient case-control cluster sampling designs for population-based epidemiological studies. In: Spatial and Spatio-temporal Epidemiology, vol. 26, pp. 95–105. Elsevier (2018). doi: 10.1016/j.sste.2018.05.002
17. Subbotin, S., Oliinyk, A.: The Sample and Instance Selection for Data Dimensionality Reduction. In: Szewczyk R., Kaliczyńska M. (eds.) Recent Advances in Systems, Control and Information Technology. SCIT 2016. Advances in Intelligent Systems and Computing, vol. 543, pp. 97-103. Springer, Cham (2017). doi: 10.1007/978-3-319-48923-0_13
18. Subbotin, S.: The Instance and Feature Selection for Neural Network Based Diagnosis of Chronic Obstructive Bronchitis. In: Bris R., Majernik J., Pancerz K., Zaitseva E. (eds.) Applications of Computational Intelligence in Biomedical Technology. Studies in Computational Intelligence, vol. 606, pp. 215-228. Springer, Cham (2016). doi: 10.1007/978-3-319-19147-8_13
19. Elliott, M.R., Valliant, R.: Inference for Nonprobability Samples. In: Statistical Science, vol. 32(2), pp. 249–264. Project Euclid (2017). doi: 10.1214/16-STS598
20. Etikan, I., Musa, S.A., Alkassim, R.S.: Comparison of Convenience Sampling and Purposive Sampling. In: American Journal of Theoretical and Applied Statistics, vol. 5(1), pp. 1–4. Science PG, New York (2016). doi: 10.11648/j.ajtas.20160501.11
21. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), last accessed 2019/01/10
22. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/Activity+ recognition+with+healthy+older+people+using+a+batteryless+wearable+sensor, last accessed 2019/01/10
23. Ravindran, A., Ragsdell, K.M., Reklaitis, G.V.: Engineering optimization: methods and applications. John Wiley & Sons, New Jersey (2018)
24. Subbotin, S.: Quasi-Relief Method of Informative Features Selection for Classification. In: 2018 IEEE 13th International Scientific and Technical Conference on CSIT, pp. 318-321. IEEE Press, Lviv, Ukraine (2018). doi: 10.1109/STC-CSIT.2018.8526627