DOI: 10.15439/2016F122
ACSIS, Vol. 8. ISSN 2300-5963

Utilizing an ensemble of SVMs with GMM voting-based mechanism in predicting dangerous seismic events in active coal mines

Łukasz Podlodowski
National Information Processing Institute
al. Niepodległości 188b 00-608 Warsaw, Poland
Email: lukasz.podlodowski@gmail.com

***Abstract*—This paper presents an application of a Gaussian Mixture Model-based voting mechanism for an ensemble of Support Vector Machines (SVMs) to the problem of predicting dangerous seismic events in active coal mines. The author proposes a method of preparing an ensemble of SVMs with different parameters and using the "wisdom of the crowd" for a classification problem. Experiments performed during the research showed an improvement in the quality of the classification after the mixture of Gaussian distributions was applied as votes distribution. The author also proposes a method of data selection for long sequences of measurement arranged chronologically with highly unbalanced occurrence of the positive class in the two-class classification problem. Finally, using the proposed model to solve the problem defined by the organizers of AAIA'16 DM showed an increase in the stability of the ensemble classifier and an improvement in the quality of the classification problem solution.**

I. Introduction

THE AIM of this paper is to present a solution to the problem introduced in AAIA'16 Data Mining Challenge: Predicting Dangerous Seismic Events in Active Coal Mines [1]. The task was related to the issue of predicting periods of increased seismic activity that may cause life-threatening accidents in underground coal mines. The task was divided into a classification problem of risk states with low hazard called ”normal” and the state with high hazard called ”warning”. Application of hybrid methods of machine learning for a similar problem was presented in [2], but instead of a two-class problem, the authors of the experiments proposed a tripartite division: ”normal”, ”warning” and ”hazard”, and focused on the medium-term (several minutes) forecasting of the maximum methane concentration at the wall end area. They also pointed out that ”mathematical models correlating methane emission with methane content of a seam, ventilation method and geological features of mine workings facilitate overall prediction of average methane concentrations during exploitation of a working, nevertheless they cannot be applied for a direct short- or medium-term prediction of methane concentration” [2]. The problem discussed at AAIA'16 Data Mining Challenge focused on a different scope of time. Granulation of data is adjusted to one-hour windows. During this time, various kinds of information are accumulated, i.e. the number of registered seismic bumps of a specific energy level, or the average activity of the most active geophone. This problem does not allow to use information on the dynamics of changes within the hour during which signals were collected, and forces the participants of the challenge to process coarse-grained information about general characteristics of the signals.

Another solution for a similar problem with regression rule learning was described in [3]. The main objective of research presented in [4] was to reduce the number of forecasting errors during monitoring natural hazards and machinery in coal mines, achieved by the application of the regression rule induction, the k-nearest neighbors method, and the time series ARIMA forecasting.

A. *Proposed solution*

I propose a solution based on an ensemble of Support Vector Machines (SVM) of the kind described in [5][6], and on a voting mechanism based on the Gaussian Mixture Model described in [7]. Common approach based on ensemble of classifiers like boosting and bagging focus on preparing different training data set for each member of ensemble [8][9]. Instead of that my solution focus on receiving different information by changing parameters of SVMs in ensemble. Each SVM was trained on the same data. The GMM voting-based mechanism allows to extract correlation between SVMs outputs and evaluate likelihood of class occurrence. Process of preparing solution is illustrated in Fig. 1

The solution ranked 4th in AAIA'16 Data Minning Challenge with a final result 0.934. The final results were evaluated in accordance with Area under Curve values on a specially curated testing set. Finding parameters of model was mostly leaded by data driven strategy. Preparing solution focused on achieving balance between quality of solution scored by AUC value evaluated in cross-validation procedure and computation complexity.

II. Preparing data

The data set included 133 151 records, each corresponding to a 24-hour measurement. Vectors had 541 columns. Values stored in a single record can be divided into two separate parts. The first part consists of an identifier of the main working site and 12 other characteristics related to the whole period of 24 hours described by the record. The second part is composed of

978-83-60810-90-3/$25.00©2016, IEEE

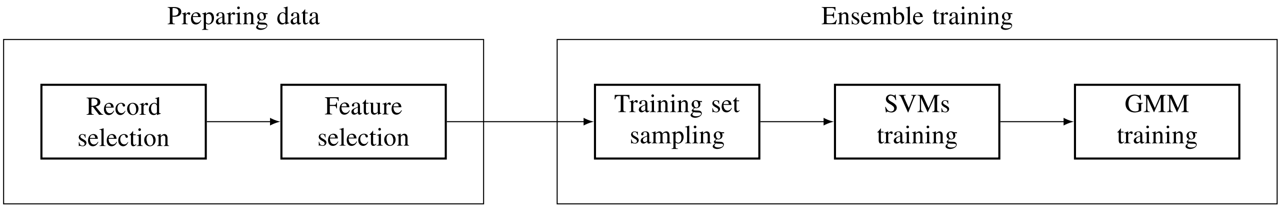

Fig. 1. Procedure of preparing data and training GMM-SVM classifier. Ensemble training was repeated 300 times and the best classifier was chosen based on cross-validation results.

hourly aggregated measurements, thus for each characteristic it includes 24 consecutive values associated with readings of geophones [1].

Measurement was labelled as ”normal” or ”warning” which indicate whether a total seismic energy perceived with 8 hours after the period covered by a data record exceeds the warning threshold, in correspondence with the classes prepared for solving the classification task. The distribution of classes was highly unbalanced. The ”normal” class covered 130 187 of all records, while the ”warning” class only corresponded to 2 962 ”warning” measurements. Records were sorted chronologically, which highlighted the tendency of ”warning” states to occur in short sequences. The testing set provided by AAIA’16 Data Mining Challenge consisted of 3 860 records.

A. Record selection

Before any operations on the data, the set vectors were linearly normalized to $[-1; 1]$. The first step towards the proposed solution was to choose the data vectors that would provide the most discriminative information. I assumed that this information was kept in boundary periods of an occurrence of ”warning” measurement. Because of that, ”normal” records were limited to the period corresponding to six hours before and to six hours after the ”warning” sequence occurred. This approach permits focusing on the most important records and to solving the problem of highly unbalanced classes occurrence in the training set. In the next part of this paper all references to the training set are corresponding to the set prepared in this way.

B. Preparing base model

The next step was to prepare a base classification model which allows future data analysis. Since the problem is a high-dimensional one, it required a model with high resistance to high dimensionality of data. Another criterion of choosing the model was the fact that almost all attributes in a vector was represented by floating or integers values and represented measurements of signal sensors which suggested that data could be naturally represented in a Hilbert space. The chosen model was SVM. The procedure of adjusting parameters of hyperplane splitting space into areas corresponding to two classes is not highly sensitive to high dimensionality of space. Moreover, expanding dimensionality of space with a constant number of data records could enable the SVM to simplify finding optimal hyperplane splitting space into two subspaces problem [8]. The SVM was also designed to solve the problem of classification of two classes. Finally, proven high effectiveness of SVM for a classification problem [10] [11] made it a natural candidate to being the base classification model for this problem. Radial Basis Function (RBF) was chosen as kernel function, for it allows SVM to map a non-linear relationship between attributes and outputs. This ability is not approachable for the linear kernel. In [12], it has been shown that the linear kernel is a special case of the RBF kernel.

Before proceeding further, the SVM was trained on a set with all attributes. Parameters of the SVM were adjusted based on the grid search procedure. In the end, the base SVM took parameters $\gamma = 0.015625$ and $C = 2$.

C. Feature selection

Firstly, record attribute corresponding to the record ID was excluded from the training set. 529 attributes of training set records was grouped in a sequence of 24 elements corresponding to values in 24-hour period of measurement. A backward elimination was performed to drop out some of these sequences. The evaluation was based on 2-fold cross-validation method. The influence of removing the whole 24-hours sequence was investigated on each step. The procedure showed that removing 12 sequences significantly increased the accuracy of the base classificator and limited vector to 252 attributes. Four attributes corresponding to the latest available hazard assessment prepared by experts took on ordinal values, which represented the level of hazard were transformed to choose only from two values ”no hazard” or ”hazard”.

III. Classification model

After preparing the data, a classifier based on the generated ensemble of SVMs classifier was proposed. Creation of this ensemble could be divided into two steps: choosing appropriate SVMs classifiers and preparing the voting mechanism.

A. Choosing SVMs

At this step, a grid search was performed and the best result it yielded was collected. I decided to use the base SVM described in the previous section of this paper and two additional SVM classifiers.

Since the high value of C allows the SVM to select more vectors as support vectors, the method could overfit. Thus, I took the first additional SVM $\gamma = 0.015625$ and $C = 1$. A second additional SVM was selected by getting a smaller value of γ, because a high value γ parameter could prevent the

SVM from finding the boundaries which allow to generalize the "shape" of the area covered by class. Based on the cross-validation results, I have chosen the SVM with parameters $\gamma = 1.22 \times 10^{(-4)}$ and $C = 1024$. Because procedure of training ensemble was repeated 300 times to avoid underfitting problem number of additionally SVMs was limited for computation time reduction.

Choosing base SVM and two additional SVMs for ensemble allow to achieve balance between computation complexity and quality of classification.

B. Voting mechanism

Instead of a simple voting mechanism, the GMM was used to represent a distribution of voting for each class. Since the testing set contains only 3 860 records, it was extremely probable that the sample would not have the same a priori distribution as the training set. It limits the possibility of a correct application of Bayes theorem in the classification model based on the estimated priori distribution. The priori likelihood of occurrence for all classes was assumed as equal in the testing data. In order to make the data represent the priori, I have limited the training set of the "normal" class for GMM to four hours before and four hours after the "warning" sequence occurs. The experiments showed that the results of the model has improved, since the data information about working wall, where measurement had been collected, was added to the vector. The working walls could be correctly identified by their IDs, but IDs do not fit well into the normal distribution âĂ¿ the base of the GMM. Hence, the ID was replaced with the height of a working wall.

The m parameter, representing the number of Gaussian components in the GMM for each class, was estimated on the basis of choosing the best results of the cross-validation procedure. If we describe parameters of a distribution as θ, a density function of the mixture distribution of features is described by:

$$f(x) = \sum_{i=1}^{m} w_i p(x \mid \theta), \tag{1}$$

where $p(x \mid \theta) = \mathcal{N}(x \mid \mu_i, \Sigma_i)$ corresponds to normal distribution that:

$$p(x \mid \theta) = \frac{1}{2\pi^{\frac{d}{2}}\sqrt{|\Sigma_i|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right). \tag{2}$$

Σ_i, μ_i and w_i are covariance matrix, mean vector and w_i represent the weight of the ith component in the mixture and d is a number of dimension in the modeled space. Parameters were estimated in EM procedure [13]. If y is assumed to be an unobserved data, then EM method takes the form of:

E-step: calculate the expectation of the unobserved data $E_{f(y|x,\theta^{(t)})}\left[\log f(x, y \mid \theta^{(t)})\right]$

M-step: find $\theta^{(t+1)}$ such that:
$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} E_{f(y|x,\theta^{(t)})}\left[\log f(x, y \mid \theta)\right]$

For n elements in the training set and an unobserved variable y_i^j takes:

$$y_i^j = \begin{cases} 1, & \text{if } i\text{th element was generated by } j\text{th component,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

To estimate parameters of the jth Gaussian component estimators take the form of:

$$w_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} E(y_i^j \mid x_i, \theta^{(t)});$$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n}\left(E(y_i^j \mid x_i, \theta^{(t)})x_i\right)}{\sum_{i=1}^{n} E(y_i^j \mid x_i, \theta^{(t)})}; \tag{4}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{n}\left(E(y_i^j \mid x_i, \theta^{(t)})(x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T\right)}{\sum_{i=1}^{n} E(y_i^j \mid x_i, \theta^{(t)})}.$$

The mixture of Gaussian components was prepared as a distribution of feature occurrences under the condition of class occurrence. Finally, the likelihood of the "warning" class was predicted on the basis of a posterior likelihood, which was evaluated basing on the Bayes theorem [14].

As Table I shows, the GMM with only one Gaussian component obtained the best results. It suggests that the SVMs' results tend to evaluate likelihood unanimously, which reduced the GMM-based voting mechanism to a single Gaussian component discriminant analysis.

C. Learning classifier

The training set for the ensemble of SVMs was different than for the GMM. Ensemble of SVMs was trained on a random sample of 30 percent of the training data. The objective of this mechanism was to avoid the overfitting problem.

Since the ensemble of SVMs was trained only on 30 percent of the training data, the risk of underfitting increased considerably. In order to solve this problem, the procedure of learning classifier GMM-SVMs was repeated. The model was learnt 300 times and the best train set for SVMs was chosen basing on the cross-validation results.

IV. Experimental results

Application of the SVM to the proposed solution was based on a LIBSVM implementation [15]. The organizers of AAIA Data Mining Challenge provided the training set in two steps. Initially, only half of the training data was available for participants. The rest of the training data was divided into four parts and supplied after some conditions, associated with the number of submissions, were met. My experiments focused on the quality of the classification performed by different approaches and the variance of results of SVM ensembles. For research variance of different strategy of voting I had collected results of an average voting, which corresponds to the voting based on simple mean of generated outputs of each SVM in the ensemble and a GMM-based voting mechanism times 50 and then estimated the mean and standard deviation of the results.

The experiments concerned with data preparation were performed on the training set provided in the first stage. The quality of solution was described by AUC and evaluated in cross-validation procedure. Other experiments, concerning the classifier quality, covered the whole training data. All cross-validation procedures were performed with fold = 2. As shown in the Table I, GMM-SVM provided the best results. I compared the average voting strategy with the GMM voting mechanism. SVM ensemble based on a simple average voting achieved worse results, but still better than a single SVM classifier.

As shown in Table II, the results of an average voting are not stable. One reason may be that all SVMs lacked the ability to cope with underfitting, which resulted in higher standard deviation of the AUC results. Since the GMM learning method calculates parameters to fit the best maximum likelihood of the prediction based on the training set, it has the ability to obtain information on the errors of each SVM and the correlation between their outputs. This allows for the use of information about the areas in space that were problematic for some of the ensemble classifiers.

TABLE I
EXPERIMENTAL RESULTS - CLASSIFICATION QUALITY

Data set	Experiment	AUC %
First part of training data	Raw data	69.098
	After backward elimination	71.923
Whole training data	SVM	71.346
	Average vote	73.428
	GMM-SVM m = 1	**75.346**
	GMM-SVM m = 2	74.474

TABLE II
EXPERIMENTAL RESULTS - STABILITY OF VOTING MECHANISMS

Voting mechanism	Mean AUC %	Standard deviation of AUC %
Average vote	54.537	17.055
GMM-SVM m = 1	73.746	0.746
GMM-SVM m = 2	72.963	0.831

V. Summary

This paper presents an application of the GMM-based voting mechanism for an ensemble of SVMs for the problem of predicting dangerous seismic events in active coal mines. The problem defined by the organizers of AAIA Data Mining Challenge allows for the successful use of GMM-SVM model of classification. My experiments showed that using the GMM voting instead of the average of outputs allows to decrease model variance. The GMM also makes obtaining information about classifier errors in the ensemble possible.

References

[1] Aaia'16 data mining challenge: Predicting dangerous seismic events in active coal mines. [Online]. Available: https://knowledgepit.fedcsis.org/contest/view.php?id=112

[2] M. Sikora, Z. Krzystanek, B. Bojko, and K. Śpiechowicz, "Application of a hybrid method of machine learning for description and on-line estimation of methane hazard in mine workings," *Journal of Mining Science*, vol. 47, no. 4, pp. 493–505, 2011. doi: 10.1134/S1062739147040125. [Online]. Available: http://dx.doi.org/10.1134/S1062739147040125

[3] M. Kozielski, A. Skowron, Ł. Wróbel, and M. Sikora, *Beyond Databases, Architectures and Structures: 11th International Conference, BDAS 2015, Ustroń, Poland, May 26-29, 2015, Proceedings*. Cham: Springer International Publishing, 2015, ch. Regression Rule Learning for Methane Forecasting in Coal Mines, pp. 495–504. ISBN 978-3-319-18422-7. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-18422-7_44

[4] M. Sikora and B. Sikora, "Improving prediction models applied in systems monitoring natural hazards and machinery," *International Journal of Applied Mathematics and Computer Science*, vol. Vol. 22, no. 2, pp. 477–491, 2012. doi: 10.2478/v10006-012-0036-3. [Online]. Available: http://www.degruyter.com/view/j/amcs.2012.22.issue-2/v10006-012-0036-3/v10006-012-0036-3.xml

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297. doi: 10.1007/BF00994018. [Online]. Available: http://dx.doi.org/10.1007/BF00994018

[6] V. Kecman, *Support Vector Machines: Theory and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, ch. Support Vector Machines – An Introduction, pp. 1–47. ISBN 978-3-540-32384-6. [Online]. Available: http://dx.doi.org/10.1007/10984697_1

[7] S. E. Jelali, A. Lyhyaoui, and A. R. Figueiras-Vidal, "Applying emphasized soft targets for gaussian mixture model based classification." in *IMCSIT*. IEEE, 2008. doi: 10.1109/IMCSIT.2008.4747229. ISBN 978-83-60810-14-9 pp. 131–136. [Online]. Available: http://dblp.uni-trier.de/db/conf/imcsit/imcsit2008.html#JelaliLF08

[8] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*, ser. Springer series in statistics. New York: Springer, 2009. ISBN 978-0-387-84857-0 Autres impressions : 2011 (corr.), 2013 (7e corr.).

[9] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738

[10] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, no. 1âĂ¿2, pp. 169 – 186, 2003. doi: http://dx.doi.org/10.1016/S0925-2312(03)00431-4 Support Vector Machines. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231203004314

[11] M. Ochab and W. Wajs, "Bronchopulmonary dysplasia prediction using support vector machine and libsvm," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014. doi: 10.15439/2014F111 pp. pages 201–208. [Online]. Available: http://dx.doi.org/10.15439/2014F111

[12] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, Jul. 2003. doi: 10.1162/089976603321891855. [Online]. Available: http://dx.doi.org/10.1162/089976603321891855

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 1–38, 1977. doi: 10.2307/2984875. [Online]. Available: http://web.mit.edu/6.435/www/Dempster77.pdf

[14] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Pearson Education, 2003, ch. 13. Uncertainty, pp. 466–486. ISBN 0137903952

[15] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011. doi: 10.1145/1961189.1961199. [Online]. Available: http://doi.acm.org/10.1145/1961189.1961199