Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 369–373
DOI: 10.15439/2014F497
ACSIS, Vol. 2

Identification of Key Risk Factors for the Polish State Fire Service with Cascade Step Forward Feature Selection

Piotr Płoński
Institute of Radioelectronics,
Warsaw University of Technology,
Nowowiejska 15/19,00-665 Warsaw, Poland
Email: pplonski@ire.pw.edu.pl

***Abstract*—The Polish State Fire Service gathers information about incidents which require their intervention. This information is stored to document the events. However, it can be very useful for new officers training, better identification of threats and planning of more effective procedures. The identification of key risk factors for casualties among firefighters, children or other involved people was a topic of data mining competition organized as a part of 1st Complex Events and Information Modelling workshop devoted to the fire protection engineering. The task of the competition was to find ten subsets of features for ten Naive Bayes classifiers. The ensemble output was used to predict occurence of casualities. Herein, the solution description that took 5th place is presented. The proposed method used cascade step forward feature selection procedure to find features subsets.**

***Index Terms*—key risk factors, fire service, Naive Bayes, feature selection, cascade step forward**

I. Introduction

THE POLISH EWID [2] reporting system is the Incident Data Reporting System (IDRS) used by Polish State Fire Service to gather information of their interventions in incidents. This data documents historical events. However, useful knowledge could be extracted from them, which can be later used for new officers training, preparation of safer and more effective procedures, and better understanding of danger factors in incidents [5], [4], [8]. The identification of key risk factors for casualties among firefighters, children and other people involved was a topic of data mining competition organized within the 9th International Symposium on Advances in Artificial Intelligence and Applications (AAIA) and was an integral part of the 1st Complex Events and Information Modelling workshop devoted to fire protection engineering. The competition results will bring data-driven insights into key risk factors in incidents and contribute to safety improvement, which is important for Fire Service supporting systems [6],[7].

The competition dataset comes from reports of the EWID system, which documents actions carried out by the Polish State Fire Service within the city of Warsaw and its surroundings in years 1992–2011. Each report obtains a feature vector descriptor after preprocessing [4], [5]. The competitors task was to find ten subset of features among over 11,000 discrete attributes describing 50,000 reports, which are relevant to the safety of people in incidents. Based on selected features, the ensemble of ten Naive Bayes classfiers [3] was created for each of three decisions variables:

1) injured firefighter in the action,
2) injured children in the incident,
3) other injured people involved.

They were used to evaluate the competition score metric, which considered the performance of the classfiers on each of the decision variable and penalizes large feature subsets. It is worth to note, that the same ten subsets of features were used in Naive Bayes construction for all decisions variables. The additional obstacle in analysis was sparsity of training data and rare occurance of positive values in decision variables.

The task of the competition can not be simplified to a sole feature selection problem. It is a problem of feature selection for ensemble of classifiers which should have the highest average accuracy in predicting three various dependent variables simultaneously with the smallest possible number of features. The proposed method used a cascade step forward selection of features that maximize the competition score metric on cross validation (CV) on training dataset. In each selection step, previously chosen features subsets were considered, therefore the proposed method is called 'Cascade Step Forward' (CSF) feature selection. The CSF procedure was speeded-up by initial features filtering and storing information about values occurences in CV folds.

The atricle is organized as follows: firstly detailed description of competition dataset, task and score metric are desribed; secondly, the proposed method is presented; then, obtained results are shown; finally, the conclusions and directions for future research are presented.

II. Methods

A. Data description

The training dataset available for participants consists of 50,000 incident reports. Each report was described using 11,852 discrete features. The majority of features were binary, with only few features with more distinct values (up to 5 values). The details of number of discrete values in features are

978-83-60810-58-3/$25.00

presented in Table I. The major values in the training dataset were zeros. From all 592,600,000 available values in training dataset only 5,217,892 have non zero values, which is only 0.8805% of all values. The sparsity and high dimensionality of data was implied by the nature of considered problem. The features correspond to the number of distinct words in the textual part of the reports (after lemmatization) and to several hundreds of features from the quantitative part of the reports [4], [5].

TABLE I
NUMBER OF DISCRETE VALUES IN FEATURES.

Discrete values	2	3	4	5
# of features	11826	9	7	10

For each report there were associated three binary decision varibles. The first decision attribute indicates incidents resulting in injury or death of a firefighter or a member of rescue team. The second decision variable indicates cases in which there were children among injured people and the third attribute identifies situations where civilians were hurt. All three decision attributes are highly imbalanced, since the positive classes correspond to relatively rare events. The details of positive values occurence in decision variables are presented in Table II

TABLE II
NUMBER OF POSITIVE VALUES IN DECISION VARIABLES.

Decision variable	# of positive values	Percentage
1	199	0.40%
2	366	0.73%
3	2955	5.91%

Let's denote dataset as $D = \{X_1, X_2, ..., X_N, Y_1, Y_2, Y_3\}$, where $N = 11,852$ is a feature number, X_i is a i-th feature vector and Y_1, Y_2, Y_3 stand for three decision variables, injury of firefighter, children, other involved people, respectively.

B. Task description

The competition task was to select ten subsets of features. They were used to build an ensemble of ten Naive Bayes classifiers for each decision variable. The sum of the output of classifiers ensemble was used to predict the occurence of positive values in each of decision variables. The accuracy of the selected features was computed with competition metric described below. It is worth to note, that there was a lower bound limit equal 3 for number of features in each subset.

C. Evaluation metric

The competition score metric can be expressed as:

$$score(s) = F\left(\frac{1}{3}\sum_{1}^{3} AUC_i(s) - \left(\frac{|s| - 30}{1000}\right)^2\right), \quad (1)$$

where

- $s = \{s_1, s_2, ..., s_{10}\}$ is a selected ten subsets of features,
- $|s|$ is a total number of selected features with repetitions,
- AUC_i is Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) [3] computed for i-th decision variable,
- $F(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$

The first term of eq.1 computes the average performance of classifier ensemble on all decision variables, whereas the second term penalizes the solutions with large number of selected features. It is worth to note, that penalization term vanishes when exactly three features are selected in each subset.

D. Proposed Method

The competition used a Naive Bayes classifier (NBC) [3] to evaluate the metric. The NBC is a classification method, which for a given sample $\mathbf{x} = \{x_1, ..., x_K\}$, with K features, calculates the posterior probability for all $y \in Y$, $p(Y = y|X_1 = x_1, ..., X_K = x_K)$, and assigns the class with the highest posterior probability. This can be expressed as:

$$y = \operatorname*{argmax}_{y \in Y} p(Y = y|X_1 = x_1, ..., X_K = x_K). \quad (2)$$

The posterior probability can be rewritten with Byes rule, the eq.2 becomes:

$$y = \operatorname*{argmax}_{y \in Y} \frac{p(Y = y)p(X_1 = x_1, ..., X_K = x_K|Y = y)}{p(X_1 = x_1, ..., X_K = x_K)}. \quad (3)$$

The evidence probability in denominator is the same for all classes and what is more, the NBC assumes that all features are conditionally independent given decision, thus the eq.4 can be written as:

$$y = \operatorname*{argmax}_{y \in Y} p(Y = y) \prod_{i=1}^{K} p(X_i = x_i|Y = y). \quad (4)$$

For discrete features the prior and likelihood can be computed as follows:

$$p(Y = y) = \frac{M_y}{M}, \quad (5)$$

and

$$p(X_i = x_i|Y = y) = \frac{M_{x_i,y}}{M_y}, \quad (6)$$

where

- M is total number of samples,
- M_y is number of samples with class label equal y,
- $M_{x_i,y}$ is number of samples with class label equal y and X_i feature equal to x_i.

The feature selection for single NBC can be done with greedy step forward (SF) procedure [3] with maximization of score with CV. The SF algorithm starts selection with empty subset of features $S_0 = \{\}$. Afterwards it checks the performance of the classifier with addition of each of the available features. The performance is computed on repeated (R_{cv} times) CV with drawing training and testing split for each repetition. The feature X_j which maximizes the quality metric is added to the subset, $S_1 = S_0 \cup X_j$. The whole procedure is

repeated till the required number of features L is selected or the required score value is achieved. The pseudocode of SF selection for single classifier is described in the Algorithm 1 listing.

Algorithm 1: The step forward feature selection procedure for single classifier.

input : $D = \{X_1, X_2, ..., X_N, Y_1, Y_2, Y_3\}$,
N number of available features,
L number of features to select,
R_{cv} repeats in cross validation.
output: The selected optimal subset S of features.
begin
  Set $S_0 = \{\}$
  for l *in 1 .. L* **do**
    for i *in 1 .. N* **do**
      Build a classifier H_i using as a feature subset $S_{l-1} \cup X_i$
      for c *in 1..* R_{cv} **do**
        Draw training and testing split of data
        Compute perfomance of classifier H_i on testing subset;
    Select classifier H_j with the highest average accuracy
    Set $S_l = S_{l-1} \cup X_j$

The SF procedure is applicable for selecting features for single classifier. It is inefficient for selecting features for ensemble of classifiers because for every classifier the similar subset of features will be assigned. The classifier ensemble requires a diverse subset of features for each classifier to obtain high accuracy [9]. To overcome this obstacle the 'Cascade Step Forward' feature selection procedure is proposed. The CSF algorithm, contrary to SF, searches for subsets of features for each classifier in the ensemble. It applies the SF procedure to find a subset of features for each classifier. However, in candidate feature scoring the performance is computed for ensemble instead of single classifier. The CSF procedure returns a set of feature subsets $S_{all} = \{S^1, ..., S^J\}$, where J is a number of classfiers in the ensemble. The pseudocode for CSF procedure is presented in Algorithm 2 listing.

E. Implementation Details

The greedy feature selection procedure has high computational cost. However, it can be decreased with filtering the features with low likelihood values. In feature selection only attributes with likelihood values greater than threshold value t for at least one decision variable were considered. The filtering condition can be expressed as:

$$p(X_i|Y_1) > t \vee p(X_i|Y_2) > t \vee p(X_i|Y_3) > t. \qquad (7)$$

The threshold value used was $t = 0.02$. After applying the eq.7 from inintial 11852 there remained 2333 features. The CSF procedure run only on remaining features.

Algorithm 2: The cascade step forward feature selection procedure for ensemble of classifiers.

input : $D = \{X_1, X_2, ..., X_N, Y_1, Y_2, Y_3\}$,
J number of classifiers in ensemble,
N number of available features,
L number of features to select,
R_{cv} repeats in cross validation.
output: The set of feature subsets for each classifier in ensemble $S_{all} = \{S^1, ..., S^J\}$.
begin
  Set $S_{all} = \{\}$
  for j *in 1 .. J* **do**
    Set $S_0^j = \{\}$
    for l *in 1 .. L* **do**
      for i *in 1 .. N* **do**
        Build a classifier H_i^j using as a feature subset $S_{l-1}^j \cup X_i$
        for c *in 1..* R_{cv} **do**
          Draw training and testing split of data
          Compute perfomance of ensemble of classifiers $\{H^1, .., H^{j-1}, H_i^j\}$ on testing subset
      Select classifier H_i^j with the highest average accuracy
      Set $S_l^j = S_{l-1}^j \cup X_i$
    Set $S_{all} = S_{all} \cup S^j$

In the proposed solution splitting dataset into training and testing subsets was performed many times during cross validation. Therefore, the counts of values occurences were stored in each fold to speed-up process of computing priors and likelihoods. The available dataset was splitted into $F = 500$ equally sized folds, from which $F_{tr} = 50$ and $F_{te} = 450$ were drawn for the training and testing respectively. Such uncommon partition provides a quite good matching between local CV scoring and public leaderboard score. The CV scoring was repeated $R_{cv} = 20$ times for each new feature testing. The i-th fold stores information M_y^i about samples number with class label equal y, and $M_{x_i,y}^i$ about number of samples with values equal x_i and class label y for all of considered features. Therefore, the probabilites needed for NBC construction can be computed as:

$$p(Y = y) = \frac{\sum_{i=1}^{F_{tr}} M_y^i}{F_{tr}\frac{M}{F}}, \qquad (8)$$

and

$$p(X_i = x_i|Y = y) = \frac{\sum_{i=1}^{F_{tr}} M_{x_i,y}^i}{\sum_{i=1}^{F_{tr}} M_y^i}. \qquad (9)$$

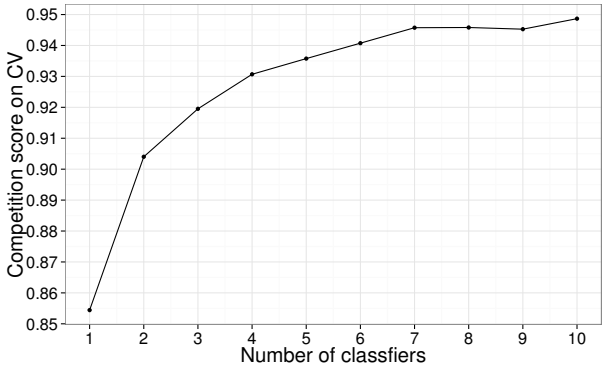

Fig. 1. The local CV score obtained for classifier ensemble in CSF feature selection.

The CSF feature selection was implemented in C++ to achieve high speed of computations.

III. Results

To omit the penalization term in the score metric (eq.1) there were selected exactly three features for each classifier. The selected features for each classfier are presented in Table III. It is worth to note, that selected features are only 0.25% of all available features. The obtained local CV scores during CSF selection for ensemble with different number of classifiers are presented in the Fig.1. It can be observed that the score is increasing when adding up to 7 classifiers into ensemble. For greater number of classifiers in the ensemble the score is stable. The local CV score was 0.9487, the public leaderboard score computed on approximately 10% of testing data was 0.9376, whereas score computed on full testing set was 0.9540. The solution that scored the 1st place achieved 0.9623 on full testing dataset, so there is only 0.0083 difference between proposed solution and the best one. The dependency between scores computed on public leaderboard and full testing dataset for solutions of all participants, with score on full testing dataset greater than 0.9, are presented in the Fig.2. It can be observed that for almost all solutions the score on public leaderboard was lowered with respect to score on the full testing dataset.

TABLE III
SELECTED ATTRIBUTES FOR EACH CLASSIFIER.

Classifier	Attribute 1	Attribute 2	Attribute 3
1	11701	5270	675
2	143	142	2182
3	691	5909	3735
4	10446	3492	2924
5	2887	8853	8914
6	7980	7148	72
7	11463	10882	1509
8	3963	258	4313
9	3596	8872	8249
10	7755	5270	6534

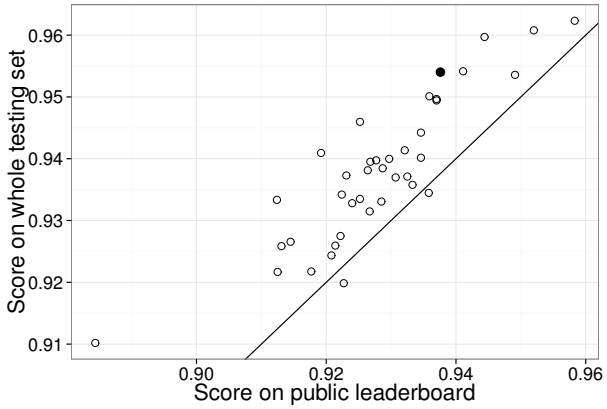

Fig. 2. The dependency between scores computed on public leaderboard (10% of testing set) and whole testing set for solutions of all participants with score on full testing set greater than 0.9. The solution presented in this paper is marked as filled black circle.

IV. Conclusion

The solution description that took 5th place in AAIA'14 Data Mining Competition: "Key risk factors for Polish State Fire Service" was presented. The proposed solution used a cascade step forward feature selection to select feature subsets for classifiers in the ensemble. The CSF maximize the competition score on cross validated training dataset in each step. To speed-up the selection process the initial filtering out of features with low likelihood were performed and number of occurence of feature values and class labels were stored in folds of training dataset. The identified key risk factors can be useful for Polish State Fire Service in new officers training, preparation of safer and more effective procedures and awareness of threats in actions.

The proposed CSF method can be applied for feature selection for other domains, for example in neuroimaging data analysis where data sets are highly dimensional and only small fraction of features are usable [10]. The performance of CSF procedure can be improved by cosidering several best features in each step instead of just one.

Acknowledgement

The author has been supported by the European Union in the framework of European Social Fund through the Warsaw University of Technology Development Programme.

References

[1] K. Bąk, A. Krasuski, M. Szczuka, "Searching for Concepts in Natural Language Part of Fire Service Reports," In: Proceedings of Concurrency, Specification and Programming; XXIII-th International Workshop, CS&P 2013, Warsaw, Poland, September 25-27, 2013.

[2] Collective Work (2001) Ewidencja zdarzen%■EWID99. Technical report, Abacus. http://www.ewid.pl/?set=rozw_ewid&gr=roz. Accessed date 23 April 2007

[3] T. Hastie, J. Friedman, R. Tibshirani, "The elements of statistical learning," Springer, 2009, DOI: 10.1007/978-0-387-84858-7

[4] A. Janusz, A. Krasuski, M. Szczuka, "Improving Semantic Clustering of EWID Reports by Using Heterogeneous Data Types," Lecture Notes in Artificial Intelligence, vol. 8170, 2013, pp. 304-314, DOI: 10.1007/978-3-642-41218-9_33

[5] A. Krasuski, A. Janusz, “Semantic Tagging of Heterogeneous Data: Labeling Fire & Rescue Incidents with Threats,” 8th International Symposium Advances in Artificial Intelligence and Applications, 2013, pp 77-82

[6] A. Krasuski, A. Jankowski, A. Skowron, D. ́lęzak, “From Sensory Data to Decision Making: A Perspective on Supporting a Fire Commander,” Web Intelligence/IAT Workshops, 2013, pp 229-236, DOI: 10.1109/WI-IAT.2013.188

[7] A. Krasuski, K. Kreński, S. Łazowy, “A Method for Estimating the Efficiency of Commanding in the State Fire Service of Poland,” Fire Technology vol.48, 2012, pp 795-805, DOI: 10.1007/s10694-011-0244-7

[8] A. Krasuski, P. Wasilewski, “The Detection of Outlying Fire Service’s Reports. The FCA Driven Analytics,” In Processings of the 11-th International Conferene on Formal Concept Analysis, 2013, pp 35-50

[9] B. Krawczyk, G. Schaefer, “A hybrid classifier committee for analysing asymmetry features in breast thermograms,” Applied Soft Computing, vol. 20, 2014, pp 112–118, DOI: 10.1016/j.asoc.2013.11.011

[10] P. Płoński, W. Gradkowski, K. Jednoróg, A. Marchewka, P. Bogorodzki, “Dealing with heterogeneous multi-site neuroimaging data sets: a study on discrimination of children dyslexia,” In: Ślęak, D., et al. (Eds.) Brain Informatics and Health, 2014, Lecture Notes in Artificial Intelligence, vol. 8609, 2014, pp 471-480