ZESZYTY NAUKOWE POLITECHNIKI BIAŁOSTOCKIEJ 2009
Informatyka – Zeszyt 4

Tomasz Łukaszuk[1]

FEATURE SELECTION USING CPL CRITERION FUNCTIONS

Abstract: Dimensionality reduction of a feature set is a common preprocessing step used for pattern recognition and classification applications. It is particularly important when a small number of cases is represented in a highly dimensional feature space. The method of the feature selection based on minimisation of a special criterion function (convex and piecewise-linear - CPL) is considered in the article. A comparison of the experimental results of this method with the results of NIPS2003 Feature Selection Challenge participant's methods is also included.

Keywords: feature selection, CPL criterion function, NIPS2003 Feature Selection Challenge

1. Introduction

The feature selection is the technique, commonly used in machine learning, of selecting a subset of the most important features for building robust learning models. By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of models constructed on the base of that data. In other words the feature selection means neglecting such measurements (features) which have no significant influence on the final decisions [4].

Dimensionality reduction is a preprocessing step commonly applied in pattern recognition and classification applications. It makes easier the next data analysis steps by alleviating the effect of the curse of dimensionality, enhancing generalization capability, speeding up learning process and improving model interpretability. Feature selection also helps people to acquire better understanding about their data by telling them which features are the important features.

The feature selection is particularly important when the data sets are composed of a small number of elements in a highly dimensional feature space. The situation when a small number of elements is represented in a highly dimensional feature space

[1] Faculty of Computer Science, Balystok Technical University, Białystok

(*long feature vectors*) usually leads to the linear separability of data sets [3]. The genomic data sets contain examples of the "long feature vectors".

This paper is engaged in the feature selection by minimization of a special convex and piece-wise linear (CPL) criterion function. The minimization process allows to calculate the parameters of hyperplane separated the learning sets and to find the best set of features ensured the linear separability of them at once. Moreover the goal of the paper is to make a comparison of described method expermiental results with the NIPS2003 Feature Selection Challenge participant's methods results.

2. Approaches to feature selection

Feature selection in substance is a task consists in removing irrelevant and redundant features from the initial data (features) set. Irrelevant and redundant features means features with no or minimal effect on later decisions.

There are two ways of selecting features set. One consists in making a ranking of features according to some criterion and select certain number of the best features. The other is to select a minimum subset of features without learning perfomance deterioration [6]. In the second way the quality of the whole subset is evaluated.

Important aspects connected with feature selection are models and search strategies. Typical models are filter, wrapper, and embedded. Filter methods use some own internal properties of the data to select features. Examples of the properties are feature dependence, entropy of distances between data points, redundancy. In the wrapper methods the feature selection is connected with the other data analysis technique, such as classification, clustering algorithm, regression. The accompand technicque helps with evaluation of the quality of selected features set. An embedded model of feature selection integreates the selection in model building. An example of such method is the decision tree induction algorithm. At each node a feature has to be selected. Basic search strategies applied in feature selection are forward, backward, floating, branch-and-bound and randomized strategies [6]. Besides there are a lot of modifications and improvements of them.

3. Linear separability of data sets and feature selection

Let us consider data represented as the feature vectors $\mathbf{x}_j[n] = [x_{j1},...,x_{jn}]^T$ $(j = 1,...,m)$ of the same dimensionality n or as points in the n-dimensional feature space $F[n]$. The components x_i of the vectors $\mathbf{x}_j[n]$ are called features. We are considering a situation, when the data can be a mixed (a qualitative-quantitative) type. Some

components x_{ji} of the vectors $\mathbf{x}_j[n]$ can be the binary ($x_i \in \{0,1\}$) and others the real numbers ($x_i \in \mathbf{R}^1$).

Let us take into consideration two disjoined sets C^+ and C^- composed of m feature vectors $\mathbf{x}_j$:

$$C^+ \cap C^- = \emptyset \,. \tag{1}$$

For example vectors from the first set represent patients suffered from certain disease and vectors from the second one represent patients without the disease. The *positive set* C^+ contains m^+ vectors $\mathbf{x}_j$ and the *negative set* C^- contains m^- vectors ($m = m^+ + m^-$).

We are considering the separation of the sets C^+ and C^- by the hyperplane $H(\mathbf{w},\theta)$ in the feature space $F[n]$.

$$H(\mathbf{w},\theta) = \{\mathbf{x} : \langle \mathbf{w},\mathbf{x} \rangle = \theta\} \tag{2}$$

where $\mathbf{w} = [w_1,...,w_n]^T \in \mathbf{R}^n$ is the weight vector, $\theta \in \mathbf{R}^1$ is the threshold, and $\langle \mathbf{w},\mathbf{x} \rangle$ is the inner product.

Definition 1. *The feature vector* $\mathbf{x}$ *is situated on the* positive side *of the hyperplane* $H(\mathbf{w},\theta)$ *if and only if* $\langle \mathbf{w},\mathbf{x}_j \rangle > \theta$ *and the vector* $\mathbf{x}$ *is situated on the* negative side *of* $H(\mathbf{w},\theta)$ *if and only if* $\langle \mathbf{w},\mathbf{x}_j \rangle < \theta$.

Definition 2. *The sets* C^+ *and* C^- *are* linearly separable *if and only if they can be fully separated by some hyperplane* $H(\mathbf{w},\theta)$ *(2):*

$$(\exists \mathbf{w},\theta) \quad (\forall \mathbf{x}_j \in C^+)\langle \mathbf{w},\mathbf{x}_j \rangle > \theta \quad \textbf{\textit{and}} \quad (\forall \mathbf{x}_j \in C^-)\langle \mathbf{w},\mathbf{x}_j \rangle < \theta \,. \tag{3}$$

In accordance with the relation (3), all the vectors $\mathbf{x}_j$ belonging to the set C^+ are situated on the positive side of the hyperplane $H(\mathbf{w},\theta)$ (2) and all the feature vectors $\mathbf{x}_j$ from the set C^- are situated on the negative side of this hyperplane.

The feature selection can be linked with searching of hyperplane $H(\mathbf{w},\theta)$ (2) separated the sets C^+ and C^-. It is possible to find a hyperplane in reduced feature space $F[n']$ $n' \leq n$. This fact results from the characteristic of the linear independence of the feature vectors $\mathbf{x}_j$ constituting the sets C^+ and C^- [2].

4. Convex and piece-wise linear (*CPL*) criterion function $\Phi_\lambda(\mathbf{w},\theta)$

If the sets C^+ and C^- are linearly separable there are very many hyperplanes $H(\mathbf{w},\theta)$ (2) divided them [3]. In order to avoiding overfitting of the model and obtainig good

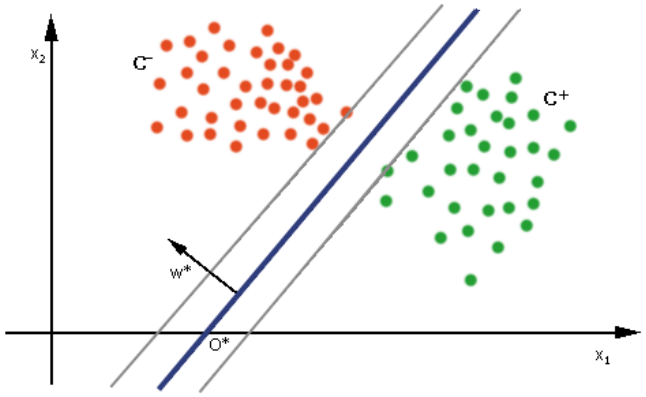

Fig. 1. Optimal hyperplane $H(\mathbf{w}^*, \theta^*)$ ensured the widest margin between itself and the elements of the sets C^+ and C^-

its generalization ability the optimal hyperplane $H(\mathbf{w}^*, \theta^*)$ should be found. Optimal hyperplane means the hyperplane ensured the widest margin between itself and the elements of the sets C^+ and C^-.

The hyperplane $H(\mathbf{w}^*, \theta^*)$ could be calculated by the minimization of the criterion function $\Phi_\lambda(\mathbf{w}, \theta)$ [3].

$$\Phi_\lambda(\mathbf{w}, \theta) = \sum_{\mathbf{x}_j \in C^+} \alpha_j \varphi_j^+(\mathbf{w}, \theta) + \sum_{\mathbf{x}_j \in C^-} \alpha_j \varphi_j^-(\mathbf{w}, \theta) + \lambda \sum_{i \in I} \gamma_i \phi_i(\mathbf{w}, \theta) \tag{4}$$

where $\alpha_j \geq 0$, $\lambda \geq 0$, $\gamma_i > 0$, $I = \{1, ..., n+1\}$.
The nonnegative parameters α_j determine relative importance (*price*) of particular feature vectors $\mathbf{x}_j$. The parameters γ_i. represent the *costs* of particular features x_i.

The function $\Phi_\lambda(\mathbf{w}, \theta)$ is the sum of the penalty functions $\varphi_j^+(\mathbf{w}, \theta)$ or $\varphi_j^-(\mathbf{w}, \theta)$ and $\phi_i(\mathbf{w}, \theta)$. The functions $\varphi_j^+(\mathbf{w}, \theta)$ are defined on the feature vectors $\mathbf{x}_j$ from the set C^+. Similarly $\varphi_j^-(\mathbf{w}, \theta)$ are based on the elements $\mathbf{x}_j$ of the set C^-.

$$(\forall \mathbf{x}_j \in C^+) \quad \varphi_j^+(\mathbf{w}, \theta) = \begin{cases} 1 + \theta - \langle \mathbf{w}, \mathbf{x}_j \rangle & if \ \langle \mathbf{w}, \mathbf{x}_j \rangle < 1 + \theta \\ 0 & if \ \langle \mathbf{w}, \mathbf{x}_j \rangle \geq 1 + \theta \end{cases} \tag{5}$$

and

$$(\forall \mathbf{x}_j \in C^-) \quad \varphi_j^-(\mathbf{w}, \theta) = \begin{cases} 1 + \theta + \langle \mathbf{w}, \mathbf{x}_j \rangle & if \ \langle \mathbf{w}, \mathbf{x}_j \rangle > -1 + \theta \\ 0 & if \ \langle \mathbf{w}, \mathbf{x}_j \rangle \leq -1 + \theta \end{cases} \tag{6}$$

The penalty functions $\phi_i(\mathbf{w}, \theta)$ are related to particular features x_i.

$$\phi_i(\mathbf{w}, \theta) = \begin{cases} |w_i| & if \ 1 \leq i \leq n \\ |\theta| & if \ \ i = n+1 \end{cases} \tag{7}$$

The criterion function $\Phi_\lambda(\mathbf{w},\theta)$ (4) is the convex and piecewise linear (*CPL*) function as the sum of the *CPL* penalty functions $\varphi_j^+(\mathbf{w},\theta)$ (6), $\varphi_j^-(\mathbf{w},\theta)$ (7) and $\phi_i(\mathbf{w},\theta)$ (7). The basis exchange algorithm allows to find the minimum efficiently, even in the case of large multidimensional data sets C^+ and C^- [1].

$$\Phi_\lambda^* = \Phi_\lambda(\mathbf{w}^*,\theta^*) = min\,\Phi_\lambda(\mathbf{w},\theta) \geq 0 \tag{8}$$

The vector of parameters w^* and the parameter θ^* define the hyperplane $H(\mathbf{w}^*,\theta^*)$. It is the best hyperplane separated the sets C^+ and C^- according to the interpretation showed on the figure 1.

5. NIPS2003 Feature Selection Challenge

NIPS is the acronym of Neural Information Processing Systems. It is the annual conference name taken place in Vancouver, Canada from 1987. Its topics span a wide range of subjects including neuroscience, learning algorithms and theory, bioinformatics, image processing, and data mining [7].

In 2003 within the framework of NIPS Conference took place the challenge in feature selection. The organizers provided participants with five datasets from different application domains and called for classification results using a minimal number of features. All datasets are two-class classification problems. The data were split into three subsets: a training set, a validation set, and a test set. All three subsets were made available at the beginning of the benchmark, on September 8, 2003. The class labels for the validation set and the test set were withheld. The identity of the datasets and of the features (some of which, called probes, were random features artificially generated) were kept secret. The participants could submit prediction results on the validation set and get their performance results and ranking on-line for a period of 12 weeks. On December 1st, 2003, the participants had to turn in their results on the test set. The validation set labels were released at that time. On December 8th, 2003, the participants could make submissions of test set predictions, after having trained on both the training and the validation set [8]. Performance was assessed using several metrics, such as:

- the balanced error rate (the average of the error rate of the positive class and the error rate of the negative class),
- area under the ROC curve (the ROC curve is obtained by varying a threshold on the discriminant values (outputs) of the classifier, The curve represents the fraction of true positive as a function of the fraction of false negative),
- fraction of features selected,

– fraction of probes (random artificially generated features) found in the feature set selected.

The NIPS 2003 challenge on feature selection is over, but the website of the challenge is still open for post-challenge submissions. One can compare results by his own method with the challenge participant's methods results.

6. Numerical experiments

There was performed the experiments with the use of earlier described feature selection methods. The input data were taken from the NIPS2003 Feature Selection Challenge web site [8]. The author's own implementation of the feature selection methods was applied during experiments. The results were formated according to directions of the challenge organizers and compared with the results of challange participans.

6.1 Data sets

There are five datasets spaned a variety of domains (cancer prediction from mass-spectrometry data, handwritten digit recognition, text classification, and prediction of molecular activity). One dataset is artificial. All problems are two-class classification problems. For the purpose of challange the data sets were prepared appropriately. Preparing the data included, amoung other things:

– preprocessing data to obtain features in the same numerical range (0 to 999 for continuous data and 0/1 for binary data),
– adding „random" features (probes) distributed similarly to the real features in order to rank algorithms according to their ability to filter out irrelevant features,
– splitting the data into training, validation and test set.

6.2 Course of experiments

Large data dimensions and connected with it a big size of data files determine using a special methods of dealing with data and strong enough computer system. The author's own implementation of the feature selection method relied on the minimization of CPL criterion function (described in section 4.) with the use of the basis exchange algorithm [1] was applied. The calculation executed on the computer equiped with 64 bit linux operation system, Intel Core2 Quad processor and 4MB RAM memory.

Table 1. The data sets of NIPS2003 Feature Selection Challenge

Dataset	Domain	Type	Features	Probes	Training exam.	Validat. exam.	Test exam.
Ascene	mass-spectrometric data, patients with cancer (ovarian or prostate cancer), and healthy	Dense	10000	3000	100	100	700
Gisette	handwritten digits: the four and the nine	Dense	5000	2500	6000	1000	6500
Dexter	texts about "corporate acquisitions"	Sparse integer	20000	10053	300	300	2000
Dorothea	discovery drugs, predict which compounds bind to Thrombin	Sparse binary	100000	50000	800	350	800
Madelon	artificial data, XOR problem	Dense	500	480	2000	600	1800

The learning data sets were constructed from the objects from training and validation sets. It means the author as a participant of the challenge starts from the second challange stage, when labels of the object from validation set are known.

The applied feature selection method generates indexes of selected features and coefficients of the hyperplane separated the learnig sets C^+ and C^- (1) corresponding to the indexes. The obtained results needed to additional process in order to submit them via NIPS2003 challenge web site and compare author's methods with the challenge entries. The results on each dataset should be formatted in ASCII files. In the separated files should be placed the classifier outputs for training, validation and test examples, the forth file should include a list of feature indexes used for classification. The author's results were formated according to the instructions.

6.3 Results

Table 2 includes the outcomes of applying author's feature selection method with the data sets of NIPS2003 Feature Selection Challenge. The numbers were received from the challenge web site after submitting formated results.

The series of results consist of the Balanced Error and Area Under Curve values (described in section 5.) defined seperately for train, validation and test sets, the number of features used by classifier and its proportion to the whole set of features and also the number of artificial features (probes) in the selected features set and its proportion to the number of all features selected by method. Besides the results concerning particular data sets, the table 2 includes the average results for all five data sets in the last row.

A perfect feature selection method should characterized by as small as possible values of Balanced Error with reference to all part of data set: train, validation and test

Table 2. Results of author's method in the NIPS2003 challenge

Dataset	Balanced Error			Area Under Curve			Features		Probes	
	Train	Valid	Test	Train	Valid	Test	#	%	#	%
arcene	0.0000	0.0000	0.3084	1.0000	1.0000	0.6916	32	0.32	18	56.25
gisette	0.0000	0.0000	0.0571	1.0000	1.0000	0.9429	222	4.44	160	72.07
dexter	0.0000	0.0000	0.1560	1.0000	1.0000	0.8440	56	0.28	12	21.43
dorothea	0.0000	0.0000	0.3401	1.0000	1.0000	0.6599	44	0.04	33	75.00
madelon	0.2665	0.2517	0.4744	0.7335	0.7483	0.5256	500	100.00	480	96.00
overall	0.0533	0.0503	0.2672	0.9467	0.9497	0.7328		21.02		64.15

(and connected with it close to 1 value of Area Under Curve). The number of features selected by a perfect feature selection method does not defined and it depends on a type of data set. Nevertheless the number should not be excessive numerous. In case of artificial features a perfect feature selection method should not choose them. It means the number of them should be as small as possible or equal to 0 in perfect situation.

The table of results shows the author's feature selection mehtod (especially its applied version) is not a perfect one. Particularly bad results were produced in case of madelon data set. Moreover the method selected too many arificial features. It concerns all data sets. The number of features used in classification are in the adequate level (except madelon data set). In case of values of Balanced Error and Area Under Curve with reference to train and validatin data sets the method turned out to be a perfect one (except madelon data set).

6.4 Conclusions

The applied method of feature selection uses the linear classifier. The properties of that kind of classifier disqualify it as a good one with xor problem. So disadvantageous results for madelon dataset follow from the character of used classifier, because medalon dataset represents exactly xor problem.

On the basis of the results the applied method is found as the method with a tendency to overfitting. The classification errors do not occure in case of train and validation data sets, whereas in the test data sets the errors are equal about 25%. The substantial number of the artificial features in the selected features set points at the tendency to overfitting, too.

6.5 Comparison with NIPS2003 Feature Selection Challenge participant's results

According to rating placed on the web site [8] and created on the basis of the results provided by the participants of NIPS2003 challenge, the author's method has placed itself on 185th position (when classification criterion is the average Balanced Error on the test set). It means about a half of the list.

A large negative influence on the rating position has very bad outcome obtained with madelon data set. The applied method does not manage with that kind of data from its nature. If the results of madelon dataset were not taken into account and the average value based on the remaining four datasets, the author's method would achieve Balanced Error value refered to the test set equal 0.2154. It would improve the rating position of method to 154th position.

The author's method occupies 134th position if the proportion of the number of features used by classifier to the number of all features is the rating criterion. The result attained for madelon dataset has again a disadvantageous influence to the situation. If the results of madelon dataset were not taken into account, the method would rate on the high enought 40th position.

The bast methods employed by participants of the challenge have obtained the Balanced Error on the test set less than 7%. However a different methods have been used with particular data sets by a single participant as a rule. For example one of the competitor describes his method in the following manner: "Combination of Bayesian neural networks and classification based on Bayesian clustering with a Dirichlet diffusion tree model. A Dirichlet diffusion tree method is used for Arcene. Bayesian neural networks (as in BayesNN-large) are used for Gisette, Dexter, and Dorothea. For Madelon, the class probabilities from a Bayesian neural network and from a Dirichlet diffusion tree method are averaged, then thresholded to produce predictions." [8]. The other participants with well results applied among other things random forests method, SVM, a different form of neural networks.

In sum it could be ascertained that the results of author's method are not extremely well. Nevertheless they are not bad, too.

7. Concluding remarks and future work

The paper presents basic assumptions of the method of feature selection based on the minimisation of the CPL criterion function. It contains the results obtained from applying of described method with the data provided by the NIPS2003 Feature Selection Challenge organizers. In comparison with participants of the challenge the

author's method has placed itself in the middle of list, particularly in case of the Balanced Error for test set, the most important rating criterion.

It needs to be noticed described experiment was the first experience of author with NIPS2003 challenge. On the basis of observations of the results list it can be stated that the approach to the challenge several times is the rule among the challenge participants. On every next approach the participant attains better rating position as the effect of improving own method. The purpose of author is also the improving his feature selection method and reaching higher rating position in the future.

The NIPS2003 Feature Selection Challenge is a good benchmark allowed on a competent estimating of the efficiency of improvements introduced in own feature selection methods as well as a comparising of own solutions with other from the research domain.

References

[1] Bobrowski L.: Design of Piecewise Linear Classifiers from Formal Neurons by Some Basis Exchange Technique, pp. 863–870 in: Pattern Recognition, 24(9), 1991.

[2] Bobrowski L., Łukaszuk T.: Selection of the linearly separable feature subsets, pp.544-549 in: Artificial intelligence and soft computing: ICAISC'2004, eds. Leszek Rutkowski, Jörg Siekmann, Ryszard Tadusiewicz, Lotfi A. Zadeh, Lecture Notes in Computer Science, vol.3070, 2004.

[3] Bobrowski L.: Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych, Wyd. Politechniki Białostockiej, Białystok, 2005.

[4] Duda O.R., Heart P.E., Stork D.G.: Pattern Classification, Second edition, John Wiley & Sons, 2001.

[5] Fukunaga K.: Statistical Pattern Recognition, Academic Press, Inc., San Diego, 1990.

[6] Liu H., Motoda H.: Computational methods of feature selection, Chapman & Hall/CRC data mining and knowledge discovery series, Chapman & Hall/CRC, 2008.

[7] `http://nips.cc/`

[8] `http://www.nipsfsc.ecs.soton.ac.uk/`

SELEKCJA CECH Z WYKORZYSTANIEM FUNKCJI KRYTERIALNYCH TYPU CPL

Streszczenie Redukcja wymiarowości zbioru cech jest często używanym wstępnym krokiem przetwarzania danych stosowanym przy rozpoznawaniu wzorców i klasyfikacji. Jest ona szczególnie istotna kiedy mała liczba obserwacji jest reprezentowana w wysoko wymiarowej przestrzeni cech. W artykule rozważana jest metoda selekcji cech opierająca się na minimalizacji specjalnej funkcji kryterialnej (wypukłej i odcinkowo-liniowej - CPL). Załączono także porównanie wyników eksperymentów uzyskanych za pomocą opisanej metody z wynikami metod uczestników konkursu NIPS2003 Feature Selection Challenge.

Słowa kluczowe: selekcja cech, funckcja kryterialna typu CPL, konkurs NIPS2003 Feature Selection Challenge

Artykuł zrealizowano częściowo w ramach grantu MNiSW 3T11F01130 i w ramach pracy badawczej S/WI/2/08 Politechniki Białostockiej.