GAME MODEL UTILIZATION FOR FEATURE RANKING AND FEATURE SELECTION

Aleš Pilný[1], **Pavel Kordík**[2], **Miroslav Šnorek**[1]

[1]Czech Technical University, Faculty of Electrical Engineering
Dept. of Computer Science and Engineering
Karlovo nám. 13, 121 35 Praha 2, Czech Republic
[2]Czech Technical University, Faculty of Information Technology
Kolejni 550/2 160 00 Prague 6, Czech Republic
pilnyale@fel.cvut.cz(Aleš Pilný)

Abstract

Majority of feature ranking and feature selection methods are designed for categorial data only and utilize statistical measures in order to rank or select features. Some of them can be used or modified for regression problems too. In this paper we present a different approach for feature ranking based on analysis of a model produced from data of interest. The main advantage of this approach is that the data mining algorithm (GAME) produces models for numerical data as well as it can be applied to categorial data. Therefore we are able to compute feature ranks for both categorial and regression problems without output discretization, which is often problematic. In this work we extract the ranking from the model topology by using statistical measures. In contrast to previous work, the rank of each feature selected by model is now computed by processing the mutual information (instead of the correlation measure) of outputs between neighboring model's neurons. The results of ranking methods were obtained from tests on artificial data sets and on well known real world data set. Our methods produce ranking consistent an in almost all cases better than in recently published studies. As an advantage these methods are applicable for numeric and categorial data as well.

Keywords: Feature Ranking, GAME, Mutual Information, Fuzzy logic, Certainty factor

Presenting Author's Biography

Aleš Pilný. After the master degree (in 2008) as a Ph.D. student joined the Computational Intelligence Group (CIG) at the Department of Computer Science and Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic. His research is focused on Feature Ranking and Feature Selection methods and its application in Data Mining. Currently he passed the doctoral exam and is working on the dissertation thesis. Author is also a member of research team of the internal grant on CTU Prague "Improvement of data preprocessing module in FAKE GAME project".

1 Introduction

The accuracy, exact time and the success of data mining generally heavily depends on quality of input features. For some problems, input features do not contain enough information to be able to perform desired task (e.g., build accurate model or classifier). There are often several possible input features that can be collected, however most of them can turn out useless. It is known and experimentally confirmed [1] that the smaller subset of attributes is often the better (it has higher classification accuracy and/or lower error). Redundant or irrelevant attribute can deteriorate the results of data mining method. When many features are available and data records are few, some data mining methods may fail to produce good models due to the course of dimensionality.

Statistical methods based on mutual information analysis [2] are able to identify most relevant input features. Algorithms (e.g, AMIFS [3]) utilizing these methods can select a representative subset of informative non-correlated features helping to overcome the curse of dimensionality.

The main drawback of these methods is that they are primary designed for nominal (discrete) variables and classification problems. In this paper, we extend selected computational intelligence methods for feature selection and feature ranking (presented in [4]) that are applicable to numerical attributes and regression problems as well. These methods use the mutual information in a ranking process instead of the previous used measure - correlation.

At first, we would like to clarify the difference among the feature ranking, feature selection and feature extraction methods. The feature ranking process only ranks all features in correspondence to their relevance while feature selection methods create a subset of the most relevant features. This subset should provide a maximal amount of information from the original subset without any redundant or irrelevant features. Methods of feature extraction, create a subset of new features by extracting the information from the original set of features.

While feature ranking simply assigns a rank (relevance) to each feature regardless of their interrelations, feature selection solves a different problem - choose the best subset of features. Note that this subset should not contains redundant features.

Generally, it is possible to classify feature selection algorithms into filters, wrappers and embedded approaches [5]. Filters evaluate quality of selected features independently from the classification algorithm, while wrapper methods depend on a classifier to evaluate quality of selected features. Finally embedded methods [5] selects relevant features within a learning process of internal parameters (e.g. weights between layers of neural networks).

The goal of this paper is to describe usage of a new measure in methods for feature ranking where these methods ranks only features preselected by the embedded feature selection algorithm. This embedded approach is based on special type of an artificial neural network, the GAME neural network [6].

Selected methods, introduced in [4], ranks features by using different approach. All of them are based on inter-relations (measured by mutual information) inside the network. Feature ranking and selection process is always performed independently.

2 Embedded feature selection process

Embedded feature selection process is an integral part of selected feature ranking methods and needs to be briefly described. This process is implemented in the FAKE-GAME [6] tool for data mining and knowledge discovery.

2.1 GAME network

A base of the FAKE-GAME tool is the Group of Adaptive Models Evolution algorithm (GAME) producing GAME networks (data mining models). The algorithm is a modification of the Multi layered Iterative Algorithm (MIA). The MIA belongs to algorithms for inductive models construction, commonly known as Group Method of Data Handling (GMDH) [7] and uses a data set to construct a model of a complex system. Layers of units transfer input variables to the output of the network. The coefficients of units transfer functions are estimated using the data set describing the modeled system. Networks are constructed layer by layer during the learning stage. Main differences between MIA and GAME are following: maximal number of unit inputs equals to the number of layer the unit belongs to, interlayer connections are allowed, transfer function and learning algorithm of units can be of several types, an ensemble of models is generated and finally the most important improvement - a genetic algorithm is used to optimize the topology. The more detailed description about the FAKE-GAME can be found in [6].

2.2 Feature selection process

Before feature ranking, the most significant features are selected. The GAME network is constructed by using a niching genetic algorithm - the corner stone of this selection algorithm. Niching methods [8] extend genetic algorithms to domains that require the location of multiple solutions. They promote the formation and maintenance of stable subpopulations in genetic algorithms (GAs). One of these methods is deterministic crowding [9]. The basic idea of deterministic crowding is that offspring is often most similar to parents. The parent is replaced by an offspring with higher fitness, and the most similar genotypic information. The reason why authors employ deterministic crowding instead of using just simple GA is the ability to maintain multiple subpopulations (niches) in the population. When the model is being constructed units connected to the most important input would soon dominate in the population of the first layer if one have used traditional GA. All other units connected to least important inputs would

show worse performance on the validation set and disappear from the population with exponential speed.

In inductive modeling one need also to extract and use information from least important features and therefore maintaining various niches in the population is preferred. The distance of genes is based on the phenotypic difference of units (to which inputs are connected). Each niche is thus formed by units connected to similar set of inputs. In the first layer, just one input is allowed and niches are formed by units connected to the same feature. After several epochs of GA with deterministic crowding the best individual (unit) from each niche is selected to survive in the layer of the model. The construction of the model goes on with the next layers, where niching is also important.

Finally we obtain the subset of features which are useful for solving the given problem. The fact that a feature is used (selected) means that it contains important information for output determination. Therefore only significant features are selected as inputs to the network and then one may compute the importance of each feature. Redundant and irrelevant features are eliminated in the genetic algorithm.

The GAME algorithm is also used in feature ranking method FeRaNGA [10] where ranks of selected features are derived from proportional numbers of connected individuals in genetic algorithms optimizing layers of units. Generally, the importance of feature increases by an amount of additional information to the information carried by already selected variables.

3 Mutual information based feature ranking methods

In previous section we have described the way how to create a subset of important features. When we need to know an importance of selected features as well, then we can analyze the topology of generated GAME network. The topology consists of different types of units (neurons with different transfer functions). When the network is ready, we know all outputs of all inner units (responses of neurons to input data vectors presented to the network). In our approach a rank of each feature is obtained as a relationship between this feature and the whole network output. As a measure of a relationship determination we used a mutual information (MI, [11]).

Let consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. In [11] mutual information $I(X; Y)$ is defined as the relative entropy between the join distribution and the product distribution $p(x)p(y)$:

$$I(X;Y) = \sum_{x \subset X} \sum_{y \subset Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}.$$

The selected methods use all neighbor neuron output inter-relations among the path between input feature and output of the network. From the definition of the GAME network there is possibility of more than one path between input and output of the network (see example in Fig. 1). These methods differ in a way of inter-relations processing and are described in following subsections. Selected methods are based on fuzzy logic and certainty factors.

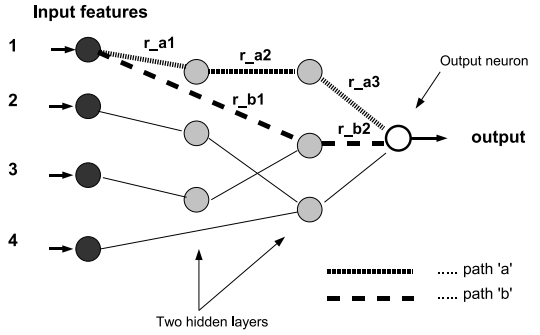

Fig. 1 Example of the final GAME network structure with four input features, one output neuron and two hidden layers. For example, path 'a' and path 'b' (highlighted) have different length. r_{a1}, r_{a2}, r_{a3} are mutual informations between neighboring neuron outputs among the path 'a' and r_{b1}, r_{b2} have the same meaning among the path 'b'.

3.1 Fuzzy Logic approach with Mutual Information - MI-FL-FR

We are finding the best relationship between input feature and output. A mutual information represents here a measure of neighboring relation. The most important relation along this path is a minimal relationship, the minimal mutual information. More than one path means also more of minima. Therefore is necessary to find the maximum of all minima among the paths between input feature and output. This process is very similar to operations from the fuzzy set theory, specially to standard complement and standard union, introduced by L. Zadeh in 1965 [12]. Therefore this approach is called Fuzzy Logic - Feature Ranking (FL-FR) with prefix MI as a mutual information measure. Computation of significance S_i for feature i can be formalised as:

$$S_i = \max(\min(r_{11}, ..., r_{1K_1}), ..., \min(r_{N1}, ..., r_{NK_N}))$$

where r_{NK_N} is inter-relation between neighbor neurons on path nr. N and K_N is K-th inter-relation on the same path.

3.2 Certainty Factor approach

In the 1980s, Dvid McAllister, developed a metric for 'certainty factors' for use in an 'expert system' (a type

of computer program)[13]. A certainty factor is used to express how accurate, truthful, or reliable one judge a predicate to be. It is one's judgment of how good the evidence is. The issue is how to combine various judgments. Let's consider a hypothesis, H, and evidence, E. The rule for evaluation is:

$$IF\ E\ is\ observed\ THEN\ H\ is\ true$$

$$(with\ certainty\ factor,\ CF = n)$$

In McAllister's scheme, a certainty factor is a number (n in the rule above) from 0.0 to 1.0 (it reflects evidence for the hypothesis only). A phrase such as 'suggestive evidence' is given a number such as 0.6; 'strongly suggestive evidence' is given a number such as 0.8. The person making the judgment uses the scale more or less as an ordinal scale. The numbers were used in a metric to permit a computer to make calculations. McAllister's rules for combining certainty factors are such that one can add new evidence to existing evidence. If the evidence is positive, this increases that certainty, as one would expect. But one never become 100 percentual certain.

In our case the certainty factor is a mutual information between neighboring neuron outputs. There were two approaches how to use the certainty factors for computing feature importance. One used a basic certainty factor judgment (chaining certainty factors). According to [4] this approach (based on a correlation measure) was significantly worse than the other approach we used in this work - combining certainty factors.

3.2.1 Combine Certainty Factors approach with Mutual Information - MI-CCF-FR

In this method certainty factors are combined along the paths and rank is assigned in dependency on maximal value of conclusion. The equation for adding two positive neighboring certainty factors (j-th and (j+1)-th) on path N is:

$$CF_{cobmi}(r_{Nj}, r_{Nj+1}) = r_{Nj} + (1 - r_{Nj}) * r_{Nj+1}$$

and importance of input feature i is then maximum of all conclusions on paths between feature i and output of the network:

$$S_i = \max(CF_{combi1}, ..., CF_{combiN})$$

where CF_{combiN} is result on N-th path.

4 Experimental data sets

We have performed various experiments on different data sets. Two artificial data sets and one real word dataset were used.

4.1 Gaussian Multivariate data Set

This artificial data set consists of two clusters of points generated from two different 10th-dimensional normal Gaussian distributions and was created by M. Tesmer and P. A. Estevez for experiments in [3]. Class 1 corresponds to points generated from N(0, 1) for each dimension and Class 2 to points generated from N(4, 1). This data set consists of 50 features and 500 samples per class. By construction, features 1-10 are equally relevant, features 11-20 are completely irrelevant and features 21-50 are highly redundant with the first ten features. Ideally, the order of selection should be: at first relevant features 1-10, then the redundant features 21-50, and finally the irrelevant features 11-20.

4.2 Uniform Hypercube Data Set

Second artificial data set consists of two clusters of points generated from two different 10th-dimensional hypercube $[0, 1]10$, with uniform distribution. The relevant feature vector (f1, f2, . . . , f10) was generated from this hypercube in decreasing order of relevance from feature 1 to 10. A parameter $\alpha = 0.5$ was defined for the relevance of the first feature and a factor $\alpha = 0.8$ for decreasing the relevance of each feature. A pattern belongs to Class 1 if ($f_i < \gamma^{i-1}$ * α / i = 1, . . . , 10), and to Class 2 otherwise. This data set consists of 50 features and 500 samples per class. By construction, features 1-10 are relevant, features 11-20 are completely irrelevant, and features 21-50 are highly redundant with first 10 features. Ideally, the order of selection should be: at first relevant features 1-10 (starting with feature 1 until feature 10 in the last position), then the redundant features 21-50, and finally the irrelevant features 11-20. This data set also come from [3].

4.3 Housing real-world data set

This Boston Housing Dataset (from ML UCI repository [14]) was taken from the StatLib library which is maintained at Carnegie Mellon University. Relevant information: Concerns housing values in suburbs of Boston, number of instances is 506 and number of attributes is 13. Attributes are continuous. Attribute Information: CRIM - per capita crime rate by town, ZN - proportion of residential land zoned for lots over 25,000 sq.ft., INDUS - proportion of non-retail business acres per town, CHAS - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), NOX - nitric oxides concentration (parts per 10 million), RM - average number of rooms per dwelling, - AGE proportion of owner-occupied units built prior to 1940, DIS - weighted distances to five Boston employment centers, RAD - index of accessibility to radial highways, TAX - full-value property-tax rate per dollars 10,000, PTRATIO - pupil-teacher ratio by town, B - proportion of blacks ratio by town, LSTAT - lower status of the population, MEDV - Median value of owner-occupied homes in dol

5 Experiments

Various experiments were performed. One on real-world data set and two on artificial data sets. First ex-

Tab. 1 Comparison on RMS error between formerly proposed methods (FL-FR and CCF-FR with correlation measure), newly modified methods (MI-FL-FR and MI-CCF-FR with mutual information measure) and ICA-FX method from [15] on real-word Housing data set (note, this is regression problem). First row indicates number of attributes selected from the subset of preselected attributes by embedded feature selection process (the smaller subset, the better attributes within). Results for ICA-FX are averages of five regression methods (MLP, SVM, 1-NN, 3-NN and 5-NN described in [15]). All methods were tested 10 times and numbers in parentheses are averages of standard deviations of 10 experiments corresponding to each regression method. The second row of each algorithm shows the best performance among the five regression methods (for ICA-FX method) or the best performance among the ten runs of FL-FR, CCF-FR, MI-FL-FR and MI-CCF-FR method. These results demonstrate that mutual information is more suitable for neuron's inter-relations measuring. The power of mutual information based approach is evident. In comparison to the correlation based methods (FL-FR and CCF-FR) MI based methods prove better results in average RMS error, smaller standard deviation and almost in every cases smaller minimum RMS error result from the second row in each experiment.

method\ # of att.	2	3	5	7	8	9	11
FL-FR	3.78 (0.08) 3.65	3.93 (0.41) 3.64	**3.15** (0.23) **2.91**	3.9 (0.32) 3.55	3.75 (0.2) 3.45	-	
MI-FL-FR	**2.87** (0.10) **2.75**	**3.54** (0.12) 3.44	**3.15** (0.09) 3.03	3.83 (0.16) 3.64	3.55 (0.11) 3.35	3.96 (0.25) 3.66	**3.24** (0.17) **3.04**
CCF-FR	5.79 (0.05) 5.71	3.98 (0.08) 3.9	4.08 (0.41) 3.79	3.48 (0.28) 3.2	4.51 (0.52) 3.761	-	
MI-CCF-FR	3.84 (0,08) 3.71	3.63 (0.12) 3.52	3.48 (0.08) 3.35	**3.01** (0.09) **2.85**	**3.25** (0.1) **3.11**	**3.22** (0.09) **3.08**	**3.24** (0.17) **3.04**
ICA-FX	-	4.09 (0.53) **3.35** (MLP)	3.74 (0.51) 3.43 (5-NN)	3.37 (0.55) 3.25 (3-NN)	-	3.48 (0.63) 3.20 (MLP)	3.61 (0.72) 3.27 (SVM)

periment (5.1) was focussed on ranking ability of proposed methods. Next two experiments (5.2) tested a stability of the attribute preselection phase during data mining model creation. All experiments have the same first step - generating of five data mining models over the data where subsets of the most significant features are selected. These subsets differ among the models because of random initialization of niching genetic algorithm (used for model creation). Configuration of this genetic algorithm was identical for all experiments (the same number of epochs and individuals, 150).

5.1 Experiment with regression data set

This experiment is designed for comparison of ranknig ability among new propsed mutual information based methods (MI-FL-FR and MI-CCF-FR), FL-FR and CCF-FR methods based on correlation measure from previous work and ICA-FX method from [15]. All these experiments were performed on real-word Housing data set in the same way.

Table 1 describes the comparison on RMS error between above mentioned methods. First row indicates number of attributes selected from the subset of preselected attributes by embedded feature selection process (the smaller subset, the better attributes are within). Results for ICA-FX are averages of five regression metohods (MLP, SVM, 1-NN, 3-NN and 5-NN described in [15]. All methods were tested 10 times and numbers in parentheses are averages of standard deviations of 10 experiments corresponding to each regression method. The second row of each algorithm shows the best performance among the five regression methods (for ICA-FX method) or the best performance among the ten runs of FL-FR, CCF-FR, MI-FL-FR and MI-CCF-FR method. These results demonstrate that mutual information is more suitable for measuring of neuron's inter-relations.

The power of mutual information based approach is evident. In comparison to the correlation based methods (FL-FR and CCF-FR) MI based methods prove better results in average RMS error, smaller standard deviation and almost in every cases smaller minimum RMS error result from the second row in each experiment. Also in comparison to the ICA-FX method MI based methods have better results in all measured parameters (RMSE, minimum RMSE and standard deviation) except minimum RMSE for No. of attributes equal to 3.

From the Table 1 is unclear which method is the best one. On the other hand one can see that MI based methods MI-FL-FR and MI-CCF-FR have in almost all cases better results than the rest of methods. It is clear that MI-FL-FR method gives better results for smaller number of selected attributes and MI-CCF-FR method gives better results for higher number of attributes. In this case (Housing data set and the Table 1) for up to 5 selected attributes is better MI-FL-FR method and for 7 and more selected attributes is better MI-CCF-FR method.

5.2 Experiments with classification data sets

In these two experiments we have tested a stability of the attribute preselection phase - embedded selection process - on artificial data sets according to results form [4]. Results on Gaussian Multivariate data set were equal in all five cases. For these five generated models embedded feature selection mechanism selects only two features. All of these selected pairs of features were equaly relevant according to the data description (only

attributes from No. 1 to 10 were selected). This step of attribute selection is not concerned by mutual information and is only an acknowledgment that selection process is working well. Mutual information was used in previous experiments where ranking ability was tested.

Second experiment in this section, done on Uniform Hypercube data set (classification problem) in the same way as first experiment, obviously showed the power of choosed approach. The selection process took into account only the most important attribute and showed us how important the selection step is. Only attribute Nr. 1 was selected. Results for these two experiments are not shown for its simplicity.

6 Conclusions

In this work we have acknowledged importance of interconnection between feature ranking methods and embedded feature selection methods. Results in embedded feature selection process from [4] were confirmed on the same artificial data sets.

Experiments with real world data set from the Table 1 also show advantages of the mutual information based approach instead of correlation based approach. There is no need to solve the problem of non-transitivity of correlation as was published in several articles. Using of mutual information brings better results in comparison to correlation based approaches.

Furthermore, mutual information based MI-CCF-FR method outperforms all results of ICA-FX approach. Interesting difference between MI-CCF-FR and MI-FL-FR methods is their opposite trend in RMS errors. The smaller subset of attributes, the better RMS error in MI-FL-FR method and the bigger number attributes, the smaller RMS error in MI-CCF-FR.

7 Acknowledgments

This research is partially supported by the research program "Transdisciplinary Research in the Area of Biomedical Engineering II" ($MSM6840770012$) sponsored by the Ministry of Education, Youth and Sports of the Czech Republic. This work is also partially supported by the internal grant of CTU Prague number SGS10/198/OHK3/2T/13.

8 References

[1] Erick Cantu-Paz. Feature subset selection, class separability, and genetic algorithms. *UCRL-CONF-202041*, 2004.

[2] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 5, NO. 4, 1994.

[3] M. Tesmer and P.A. Estevez. Amifs: adaptive feature selection by using mutual information. In *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks*, volume 1, page 308, Dept. of Electr. Eng., Chile Univ., Santiago, Chile, July 2004.

[4] Snorek M. Pilny A., Kordik P. Correlation-based feature ranking in combination with embedded feature selection. *In proceedings of International Conference on Inductive Modelling*, 2009.

[5] J. Biesiada, W. Duch, A. Kachel, K. Maczka, and S. Palucha. Feature ranking methods based on information entropy with parzen windows. pages 109–119, 2005.

[6] P. Kordík. *Fully Automated Knowledge Extraction using Group of Adaptive Models Evolution*. PhD thesis, Czech Technical University in Prague, FEE, Dep. of Comp. Sci. and Computers, FEE, CTU Prague, Czech Republic, September 2006.

[7] H. R. Madala and A.G. Ivakhnenko. *Inductive Learning Algorithm for Complex System Modelling*. CRC Press, 1994. Boca Raton.

[8] Samir W. Mahfoud. Niching methods for genetic algorithms. Technical Report 95001, Illinois Genetic Algorithms Laboratory (IlliGaL), University of Ilinios at Urbana-Champaign, May 1995.

[9] S. W. Mahfoud. A comparison of parallel and sequential niching methods. In *Sixth International Conference on Genetic Algorithms*, pages 136–143, 1995.

[10] Aleš Pilný, P. Kordík, and M. Snorek. Feature ranking derived from data mining process. *18th International Conference on Artificial Neural Networks - ICANN 2008*, pages 889–898, 2008.

[11] Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley, second edition edition, 2006.

[12] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[13] Robert J. Chassell. About certainty factors. http://www.rattlesnake.com/notions/certainty-factors.html, 2009.

[14] C. J. Merz C. L. Blake. Uci repository of machine learning databases. http://www.ics.uci.edu/ mlearn/MLSummary.html, September 2006.

[15] Nojun Kwak, Chunghoon Kim, and Hwangnam Kim. Dimensionality reduction based on ica for regression problems. *Neurocomputing*, 71:2596 2603, 2008.