Anadolu Üniversitesi Bilim ve Teknoloji Dergisi A- Uygulamalı Bilimler ve Mühendislik
Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering

2016 - Volume: 17 Number: 5
Page: 845 - 852
DOI: 10.18038/aubtda.279853
Received: 27 June 2016 Revised: 18 October 2016 Accepted: 17 November 2016

A MODIFIED T-SCORE FOR FEATURE SELECTION

Hüseyin BUDAK [1,*], Semra ERPOLAT TAŞABAT [1]

[1] Department of Statistics, Mimar Sinan Fine Arts University, Istanbul, Turkey

ABSTRACT

In this study, an alternative approach to t-score method, one of the feature selection methods, has been suggested and some analyses have been executed in order to compare t-score method and our approach. When comparing them, commonly used data sets in data mining studies, Arcene, Gisette and Madelon have been used. In line with the purpose of this study, the first 50, 100, 150 and 200 features for each data set has been selected, in consequence, 24 data subsets have been created. The classification accuracies of t-score and suggested method has been compared by using these data subsets. When calculating the classification accuracies, two commonly used methods in literature, Artificial Neural Networks and Support Vector Machines have been used. According to this study, the result of the suggested feature selection method is statistically more successful than t-score.

Keywords: Data mining, Feature selection, t-score, Classification accuracy

1. INTRODUCTION

Nowadays the developments in information technologies have given us the opportunity to build up databases in many areas and to incrementally increase the amount of data stored. The increase of this data also exceeds the expectation of the operators. In order to meet the expectations of the operators, data mining methods rise to prominence since traditional methods which analyze the big data stored in databases are inadequate. Data mining is a process comprised of not only analyzing available data but also collecting data, obtaining meaningful information from it and transforming this information to an action plan. One of the stages of this process is feature selection. Thanks to the capacity of the databases, there are hundreds, and sometimes thousands of features in the data that are used to solve real world problems. When analyzing this amount of data, the feature selection prior to multi-dimensional data analysis has become significantly important as one faces issues such as time spent on the execution, data storage expenses and the performance decrease of data mining algorithms.

The aim of this article is to study the feature selection methods and suggest a new that may be an alternative one.

2. FEATURE SELECTION METHODS

Feature selection (also known as subset selection or variable selection), can be briefly described as the selection of the best subset which can represent the original data set. Feature selection is the process of selecting the best k features among the n features in dataset by evaluating the features according to the algorithm that is used [3].

Advantages of feature selection [1]:

- reduces the dimension of the feature space and increases algorithm speed,
- removes the redundant, irrelevant or noisy data,

*Corresponding Author: huseyin.budak@hotmail.com.tr

- improves the data quality,
- transforms the data into a more understandable, simply definable and can be visualized form,
- it saves resource during data collection or during utilization,
- increases the accuracy of the resulting model.

Feature selection methods can be categorized into three groups, one of which is the filter methods (based on just statistical knowledge). Second one is the wrapper methods (perform search techniques on features). The third one is the embedded methods (based on finding the best separation criteria) [5].

In filter methods, feature selection is executed before data mining algorithm but in wrapper methods, data mining algorithm is used as a tool for feature selection. In embedded methods, data mining algorithm and feature selection are executed simultaneously.

2.1. General Steps of Feature Selection Methods

Feature Selection algorithms runs with similar steps as shown in Figure 1. Firstly, a feature subset is generated from the original data set and then an evaluation is carried out to the features discussed. As a result of the evaluation, it is decided whether the related feature is selected or not and the selected feature is added in subset. This process continues until the stopping criterion of the algorithm is achieved.

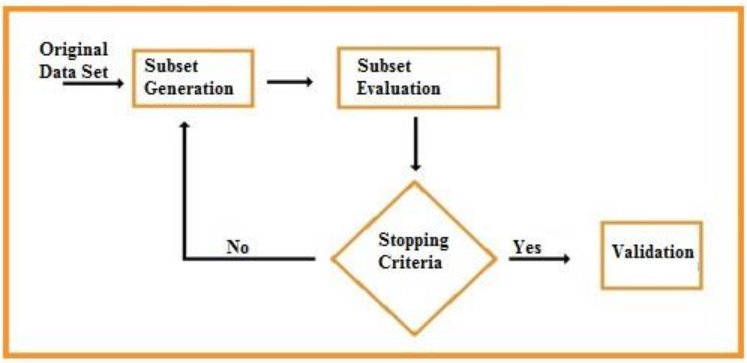

Figure 1. General feature selection flow chart [15]

2.2. Filter Methods

Filter methods, the oldest feature selection algorithms in data mining applications, selects the features based on statistical knowledge such as distance, information, dependency and consistency measures without using any classifier. Basically these methods calculate a score by using an evaluation function which runs according to the statistical measure defined for each feature in the original data set. Features that have the highest scores among calculated scores are selected into the best subset.

These methods are more suitable for big data sets as they are less complex, have less calculation cost and give faster results compared to other methods [2].

Methods;

- t-score [6]
- Fisher score [6]
- Welch t-statistics [6]

- Chi-square test [4]
- Information gain [8]
- Gain ratio [7]
- Symmetrical Uncertainty [7]
- Correlation based Feature Selection (CFS) [8]
- Relief [9]
- One-R [21]
- mRMR [10]

Since the aim of this study is to suggest an approach for t-score, detailed information related to only t-score is given in this section.

2.2.1. t-score

t-score is one of the commonly used feature selection methods. This method calculates a relation score by using the sample size, mean and standard deviation values of the features for each class. Features with less score is eliminated from the data set. The formula of t-score is shown in (1) [6].

$$t(x_i) = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sqrt{\frac{n_i^+(\sigma_i^+)^2 + n_i^-(\sigma_i^-)^2}{n_i^+ + n_i^-}}} \tag{1}$$

In this equation; + and – are class labels, μ_i^+ and μ_i^- are means of classes, σ_i^+ and σ_i^- are standard deviations of classes, n_i^+ and n_i^- are sample sizes of classes,

The feature selection process of t-score method is executed in the form of features that are ranked by descending order according to the computed scores, then desired number of features are selected starting from the top.

2.3. Wrapper Methods

Wrapper methods consist in using the prediction performance of a given learning machine to assess the relative usefulness of subsets of variables [2]. In these methods, features to be selected are determined by performing search process depending on various classification or learning algorithms that are applied on the original data set. Although wrapper methods more successful than filter methods in selecting the best feature subset, they have high calculation cost.

Methods;

- Sequential Forward Selection (SFS) [11]
- Sequential Backward Selection (SBS) [22]
- Plus l - Take Away r [13]
- Sequential Forward Floating Selection (SFFS) [12]
- Sequential Backward Floating Selection (SBFS) [12]

2.4. Embedded Methods

Because structure of the embedded methods has both classification algorithm and feature selection algorithm, classification and feature selection processes are executed simultaneously in these methods [2]. Embedded methods have higher calculation cost than filter methods just like wrapper methods.

Methods;

- Decision Tree [1]
- Support vector machines-Recursive feature elimination (SVM-RFE) [14]

3. MODIFIED T-SCORE

Suggested method is based on t-score that is the most commonly used among filter methods. As mentioned in 2.2.1, a relation score is calculated by using the sample size, mean and standard deviation values of the features for each class and features with high scores are added to subset in t-score method. In the suggested method, the researcher tried to increase the features' scores which are thought to contribute more to the success of classification. In line with this purpose, starting from the idea that if features are highly correlated with the class and low correlated with each other, they contribute more to the success of classification, and the terms r_{iy} and $\overline{r}_{ix}$ have been added to t-score formula. Since features selected by suggested method are desired to have highly correlation with the class (y) and low correlation with each other (x), r_{iy} is added to dividend of t-score formula and $\overline{r}_{ix}$ is added to denominator of the formula. Because the type of relationship is not taken into consideration when calculating the correlations for suggested method, the terms r_{iy} and $\overline{r}_{ix}$ have been used as absolute values. Thus, the formula (2) has been created. Just like in other methods, by using the suggested method, the calculated (feature) scores are listed in descending order and then features selection is executed by choosing the desired number of features starting from the top.

$$t'(x_i) = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sqrt{\frac{n_i^+(\sigma_i^+)^2 + n_i^-(\sigma_i^-)^2}{n_i^+ + n_i^-}}} \frac{\left|r_{iy}\right|}{\overline{r}_{ix}} \tag{2}$$

In this equation; $\left|r_{iy}\right|$ is the absolute correlation value of discussed feature with class labels, $\overline{r}_{ix}$ is mean of the absolute correlation value of discussed feature with the others features.

4. IMPLEMENTATION AND RESULTS

4.1. Used Data Sets and Methods

In this paper, Arcene, Gisette and Madelon data sets are used to execute experimental comparisons. These data sets are obtained from Neural Information Processing Systems Conference (NIPS) 2003 - Feature Selection Challenge [16, 17, 18, 19, 20].

First of all, scores of all features in Arcene, Gisette and Madelon data sets have been calculated for t-score and suggested feature selection method in order to analyze the data. Features in the data sets are sorted in descending order according to the computed scores and features are selected as first 50, first 100, first 150 and first 200 by starting from the top for each of the methods. Thus, 24 24 data subsets have been created. Subsequently, classification accuracies have been calculated for these data sets by applying Artificial Neural Networks and Support Vector Machines. A single hidden layer MLP (Multilayer Perceptron) and a RBF (Radial Basis Function) Kernel model (cost parameter is 10 and gamma is 0.1) have been used for classification.

In the process of classification, first, data sets have been divided into two separate groups as training set (70%) and test set (30%). Afterwards, a model has been established for training set and classification accuracies have been calculated by applying the model on the test set.

For the comparison of t-score with suggested method, t-test has been executed in order to see whether there is statistically significant difference in the calculated percentages of classification accuracy or not. In order to test the assumptions of the t-test; Kolmogorov-Smirnov test has been executed for assumption of normality and Levene test has been executed for assumption of equal variances. A significance level of 0.05 has been used in all the tests applied.

4.2. Results

In this section, graphs of the classification accuracies have been created for the purpose of comparing t-score with suggested method. The difference of classification accuracy percentages between t-score and suggested method have been demonstrated visually with graphs. Also, t-test has been executed in order to test whether this difference is statistically significant.

As shown in Figure 2, classification accuracies of suggested method in all data sets are higher than classification accuracies of t-score method. T-test analysis must be executed in order to decide whether the difference between these two methods is statistically significant or not. The data set that suitable for t-test analysis has been created by combining the classification results that specified separately according to the classification methods and data sets previously. The data set was created are given in Table 1.

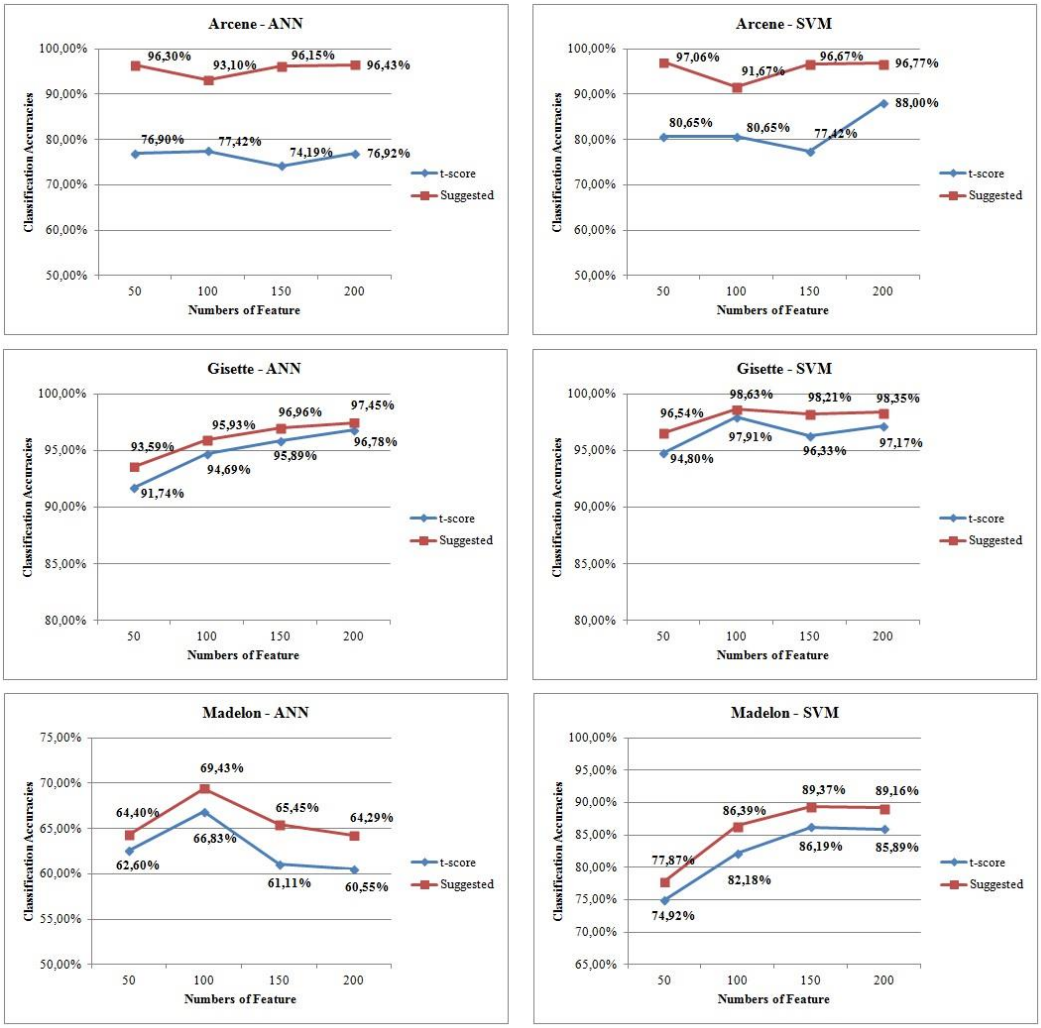

Figure 2. A comparison of t-score with suggested method by classification accuracies for Arcene, Gisette and Madelon data sets

Table 1. Classification Accuracies of T-score and Suggested Method

	t-score (%)	Suggested (%)		t-score (%)	Suggested (%)
Arcene - ANN - first50	76.92	96.30	Gisette - SVM - first50	94.80	96.54
Arcene - ANN - first100	77.42	93.10	Gisette - SVM - first100	97.91	98.63
Arcene - ANN - first150	74.19	96.15	Gisette - SVM - first150	96.33	98.21
Arcene - ANN - first200	76.92	96.43	Gisette - SVM - first200	97.17	98.35
Arcene - SVM - first50	80.65	97.06	Madelon - ANN - first50	62.60	64.40
Arcene - SVM - first100	80.65	91.67	Madelon - ANN - first100	66.83	69.43
Arcene - SVM - first150	77.42	96.67	Madelon - ANN - first150	61.11	65.45
Arcene - SVM - first200	88.00	96.77	Madelon - ANN - first200	60.55	64.29
Gisette - ANN - first50	91.74	93.59	Madelon - SVM - first50	74.92	77.87
Gisette - ANN - first100	94.69	95.93	Madelon - SVM - first100	82.18	86.39
Gisette - ANN - first150	95.89	96.96	Madelon - SVM - first150	86.19	89.37
Gisette - ANN - first200	96.78	97.45	Madelon - SVM - first200	85.89	89.16

In order to apply t-test, the data in Table 1 has been tested whether it satisfies the assumptions of t-test.

As shown in Table 2; It has been determined that the data sets of t-score and suggested methods have a normal distribution according to Kolmogorov-Smirnov test results (0.741>0.05 and 0.095>0.05) and two groups in the data set have equal variances according to Levene test result (0.677>0.05).

Table 2. The Results of T-test Assumptions

Test	P value
Kolmogorov-Smirnov (t-score)	0.741
Kolmogorov-Smirnov (suggested)	0.095
Levene	0.677

It has been determined that the data sets of t-score and suggested methods have a normal distribution according to Kolmogorov-Smirnov test results (0.741>0.05 and 0.095>0.05) and two groups in the data set have equal variances according to Levene test result (0.677>0.05).

As shown in Table 3; It has been determined that calculated classification accuracy percentages have a statistically significant difference according to t-score and suggested methods ($p=0.046 < \alpha=0.05$). When this difference is examined, it has been seen that the mean of classification accuracy percentages of suggested method (89.42) is higher than the mean of classification accuracy percentages of t-score method (82.41).

Table 3. The Result of T-test Analysis

	N	Mean	SD	T	P
t-score	24	82.41	0.120	-2.05	0.046
Suggested	24	89.42	0.117		

5. CONCLUSION

It is an inevitable fact that information technology, is in escalation and provides invaluable opportunity to handle and work with large amount of data in variety of areas. This increase in data results in surpassing the expectations of the operators. There is no doubt that traditional methods are insufficient to analyses this amount data and leads the analyzers to use data mining methods which are becoming of great importance since they meet the expectations of the operators. Therefore data mining applications come to the forefront.

One of the most important stages of data mining procedure is the process of dimension reduction. The dimension reduction is the process of removing irrelevant or redundant variables from the data set in order to resolve problems encountered in storaging big data sets and analyzing them. Feature selection is one of the most popular method among the methods of dimension reduction.

Feature selection, is described as the selection of the best subset which can represent the original data set. This process aims to reduce the number of features in the data set by selecting the most useful and important features for the discussed problem.

In this study, an alternative approach to t-score method, one of the feature selection methods, has been suggested and some analyses have been executed in order to compare them. In order to compare the feature selection methods, primarily, the graphs have been created by using classification accuracies that are calculated for all data sets and classification methods. Thus, a visual comparison of t-score with suggested method has been provided. Looking at all the graphs given, it can be concluded that classification accuracies of suggested method are higher than classification accuracies of t-score method. Moreover, it has been calculated as the mean of classification accuracy percentages of suggested method is 89.42% and the mean of classification accuracy percentages of t-score method is 82.41%. Afterwards, t-test analysis has been executed in order to test whether the difference between classification accuracy percentages of these two methods are statistically significant or not. As a result of this test, it has been determined that calculated classification accuracy percentages have a statistically significant difference according to t-score and suggested methods. Also, the average of classification accuracy percentages of suggested method are higher than the average of classification accuracy percentages of t-score method (at the 0.95 level of confidence).

Last of all, classification results obtained from the suggested method has been more successful as an alternative to t-score method, one of the most frequently used methods among filter feature selection methods. In the light of these results, it can be said that the suggested method as an alternative to t-score method can be used in feature selection performed as a preparation process for data mining applications and the suggested method can give better results.

REFERENCES

[1] Ladha L, Deepa T. Feature selection methods and algorithms, International Journal on Computer Science and Engineering 2011; 3: 1787-1797.

[2] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research 2003; 3: 1157-1182.

[3] Forman G. An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research 2003; 3: 1289–1305.

[4] SPSS Inc, SPSS Clementine 12.0 Algorithms Guide. SPSS Inc, USA: Chicago, 2007.

[5] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics, Bioinformatics 2007; 23: 2507-2517.

[6] Yıldız O, Tez M, Bilge HŞ, Akcayol MA, Güler İ. Meme kanseri sınıflandırması için gen seçimi (Gene selection for breast cancer), IEEE 20nd Signal Processing and Communications Applications Conference; 18-20 April 2012; Muğla, Turkey: IEEE, p.p 988-991.

[7] Novakavic J, Strbac P, Bulatovic D. Toward optimal feature selection using ranking methods and classification algorithms, Yugoslav Journal of Operations Research 2011; 21: 119-135.

[8] Hall MA. Correlation-based Feature Selection for Machine Learning, PhD, The University of Waikato, New Zealand: Hamilton, 1999.

[9] Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm, AAAI 1992; 2: 129-134.

[10] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology 2005; 3: 185-205.

[11] Whitney AW. A direct method of nonparametric measurement selection, IEEE Transactions on Computers 1972; 20: 1100-1103.

[12] Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection, Pattern Recognition Letters 1994; 15: 1119-1125.

[13] Stearns SD. On selecting features for pattern classifiers, 3rd International Conference on Pattern Recognition; 8-11 November 1976; Coronado, California, USA: pp 71-75.

[14] Guyon I, Weston J, Barnhil S, Vapnik V. Gene Selection for cancer classification using support vector machines, Machine Learning 2002; 46: 389-422.

[15] Liu H, Yu L. Towards integrating feature selection algorithms for classification and clustering knowledge and data engineering, IEEE Transactions on Computers 2005; 17: 491-502.

[16] http://www.nipsfsc.ecs.soton.ac.uk/datasets/ [accessed 11.01.2015]

[17] http://www.nipsfsc.ecs.soton.ac.uk/papers/NIPS2003-Datasets.pdf [accessed 11.01.2015]

[18] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, & Liotta LA, Use of proteomic patterns in serum to identify ovarian cancer, The lancet 2002; 359: 572-577.

[19] LeCun Y, Bottou L, Bengio Y, Haffner P, Gradient-based learning applied to document recognition, Proceedings of the IEEE 1998; 86: 2278-2324.

[20] Perkins S, Lacker K, Theiler J. Grafting: Fast, Incremental feature selection by gradient descent in function space, The Journal of Machine Learning Research 2003; 3: 1333-1356.

[21] Holte R C. Very simple classification rules perform well on most commonly used datasets, Machine Learning 1993; 11: 63-91.

[22] Marill T, Green D M. On the effectiveness of receptors in recognition system, IEEE Trans. Inform. Theory 1963, 9: 11-17.