A Feature Selection Technique based on Distributional Differences

Sung-Dong Kim*

Abstract: This paper presents a feature selection technique based on distributional differences for efficient machine learning. Initial training data consists of data including many features and a target value. We classified them into positive and negative data based on the target value. We then divided the range of the feature values into 10 intervals and calculated the distribution of the intervals in each positive and negative data. Then, we selected the features and the intervals of the features for which the distributional differences are over a certain threshold. Using the selected intervals and features, we could obtain the reduced training data. In the experiments, we will show that the reduced training data can reduce the training time of the neural network by about 40%, and we can obtain more profit on simulated stock trading using the trained functions as well.

Keywords: Feature Selection, Distributional Differences

1. Introduction

Features are used to specify information relevant to a certain classification problem. Generally, classification algorithms or machine learning techniques determine the relevant information using many initial features that are assumed to be relevant. Feature selection is a process of determining relevant features for a classification problem using some statistical methods. Feature selection reduces the dimensionality of the feature space and removes redundant, irrelevant, and noisy data. It plays an important role in data selection and preparation in the field of statistical pattern recognition, data mining, machine learning, and so on [1]. For example, feature selection speeds up the data mining algorithm, improves the data mining performance, and enhances the interpretability of the results.

Automatic classifications based on the trained classifiers through machine learning techniques are widely used in many areas. Machine learning requires a huge amount of training data. The data is automatically generated to include noisy data, and it is general to construct training data including many features because we don't know the relevant features for the classification problem in advance. Using too many features makes the training and the classification longer and complex, and worse still, it may cause an over-fitting problem. It is important to the efficiency and effectiveness of the machine learning to extract relevant features for the classification problem using some statistical methods, build the training data including the extracted features only, and train the classifier with the reduced training data [2]. This paper presents the feature selection technique based on distributional differences. The proposed technique selects meaningful features by the distributional differences of the features in the positive and negative data. The concept of distributional differences was also applied to the construction of automatic stock trading system [3] and the induction of stock trading rules [4].

Initial training data contains values of the many candidate features and a target value. We classify the training data into the positive and the negative data by the target value. We divide the range of the feature values into 10 intervals. For positive and negative data separately, we count the frequency of the feature values included in each interval and calculate the distributions of each interval. We select the features and the intervals of features whose distributional differences differ by more than a certain threshold value. Initial training data is reduced using the values that belong to the selected intervals, and is reduced again to include the selected features only. We discard the intervals of candidate features whose distributions in positive and negative data are similar, and remove a data with values included in such intervals. This is because we consider data with values included in such intervals as a non-distinctive one for classification. We re-build the training data using such features that have distinctive distributional differences of the feature value intervals.

This paper is organized as follows. Section 2 shows the existing studies of the feature selection methods. Section 3 describes the proposed feature selection technique and the reduction process of the initial training data. Section 4 presents the experimental results supporting the efficiency and the effectiveness of the machine learning using the reduced training data. Section 5 draws a conclusion and presents some future works.

2. Previous Works

Feature selection has been an ongoing research theme

Manuscript received February 23, 2006; accepted March 3, 2006.
This research was financially supported by Hansung University in the year 2005.

Corresponding Author: Sung-Dong Kim
* Dept. of Computer Engineering, Hansung University, Seoul, Korea (sdkim@hansung.ac.kr)

since the statistical pattern recognition in the 1970s [5], and it plays an important role in the field of machine learning and data mining [6, 7], text classification [8], and so on. There are several studies for feature selection in the field of machine learning. They can be classified by the goodness evaluation methods of the feature subset or by the existence of the class information in the data. The goodness evaluation methods of the feature subset are divided into filter model and wrapper model [9]. Feature selection for the training data with target class information is called **supervised** [10], and the contrary is an **unsupervised** feature selection [11].

In the **filter model**, irrelevant features are removed before the learning or induction process. That is, the feature selection is a pre-processing step of the machine learning. For example, FOCUS algorithm [14] and RELIEF algorithm [12] are the well-known instances of the filter model. FOCUS algorithm determines the minimum feature subset for correctly classifying the training data through the breadth-first search. Due to the consistency condition in the feature selection, satisfying that the algorithm should correctly classify the training data, FOCUS algorithm is very sensitive to noisy data or inconsistency in the training data, and is not appropriate to problems with more than 25 features. RELIEF algorithm assigns the weights representing the classification ability to the features, and selects features whose weights are bigger than the user-defined threshold value. The filter models feature the different bias in the feature selection process from that in the learning process.

The **wrapper model** evaluates the performance of the feature subset by the performance of the classifiers generated through the learning process with the initial training data that contains all candidate features [13]. It is called wrapper model because the feature selection process is wrapped by the learning process. The model has a similar bias in the feature selection process to that in the learning process, but it takes longer to select features compared to the filter model.

There have been several researches on feature selection algorithms since then. They are a kind of the filter model or the wrapper model, or hybrid methods taking advantage of both models.

3. Feature Selection Technique

We propose the feature selection technique based on distributional differences, which aims at improving the efficiency and the effectiveness of the machine learning. The technique is motivated by the fact that there are many non-effective data in the initial training data for the classification problem. We assume that such data would show similar distributional properties in positive and negative data sets. Thus, we propose the feature selection technique based on the distributional differences. The proposed technique is a kind of the filter model. It is also the supervised method because the training data contains the class information. This paper presents the feature selection technique for the two-class classification problem With the feature selection, we reduced the size of initial training data and attempted to decrease the training time and enhance the performance of the trained classifiers. Fig. 1 shows the process of the feature selection.

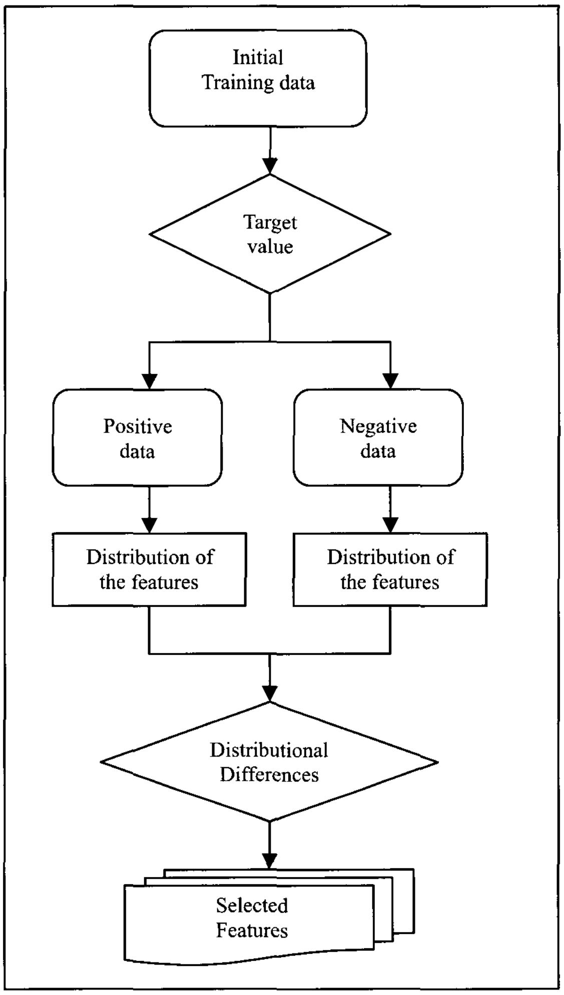
Initial Training data
Target value
Positive data
Negative data
Distribution of the features
Distribution of the features
Distributional Differences
Selected Features

Fig. 1. The overall process of the feature selection

3.1 Classification of the Training Data

The training data contains the class information for two classes. We divided the data into the positive and the negative data sets with the class information. The calculation of the distribution for feature selection was performed independently for each positive and negative data set.

3.2 Distribution Calculation of the Features

We divided the range of the feature values into 10 intervals[‡] and counted the frequency of the feature values included in each interval. Then, we calculated the

[‡] We select the number of intervals as 10 randomly.

distributions of each interval by dividing the counts by the total number of data. The following equations are for the distribution calculation in each positive and negative data set

$$p_p(\mathrm{int}_i(f^k)) = \frac{|\mathrm{int}_i(f^k)|}{N_p},$$

$$p_n(\mathrm{int}_i(f^k)) = \frac{|\mathrm{int}_i(f^k)|}{N_n},$$

where $p_p(\mathrm{int}_i(f^k))$ is the probability of the *i*th interval of the *k*th feature in the positive(*p*) data set, and $p_n(\mathrm{int}_i(f^k))$ is the one in the negative(*n*) data set. N_p is the total number of positive data, and N_n is that of the negative data.

3.3 Feature Selection

We extracted the intervals of the features whose distributional differences in positive and negative data sets were bigger than the threshold value. The selection of the intervals is based on the following selection function

$$select(\mathrm{int}_i(f^k)) = \begin{cases} 1, \text{ if} p_p(\mathrm{int}_i(f^k)) - \\ \quad p_n(\mathrm{int}_i(f^k)) > \varepsilon, \\ 0, \text{ otherwise} \end{cases}$$

where ε is the threshold value for the distributional difference.

The intervals of the features whose distributional differences in positive and negative data sets were bigger than the threshold value can be considered as the relevant ones to the classification. Therefore, those intervals were selected for re-building the training data.

3.4 Reduction of the Training Data

One data item in the initial training data set consists of values of the candidate features. The reduction of the initial training data consists of two steps. The first step removes data without the values included in the selected intervals of the features. A data is considered the necessary data to the learning if at least one value of the data belongs to the selected intervals of the features. In this step, we removed the data that were considered as not affecting the learning process. This step is called the **data filtering** step because the number of the data in the training set decreases. In the second step, we removed the features whose value intervals were not selected at all. When the number of the features in the initial training set is *n*, the feature selection of this step is represented as:

$$IFS = \{f_1, f_2, \ldots, f_n\} \Rightarrow$$
$$SFS = \{sf_1, sf_2, \ldots, sf_m\}, \text{ where } n > m.$$

In the above equation, *IFS* means the initial feature set and *SFS* means the selected feature set.

The reduction of the training data with the selected features causes the reduction of the dimensionality of the data. This step is called the **dimension reduction** step because the data dimension is reduced. Then, the initial data is reduced as follows:

$$reduce(init\_data) = (sf_1, sf_2, \ldots, sf_m, class_{init\_data}),$$
$$\text{where } init\_data = (f_1, f_2, \ldots, f_n, class_{init\_data}).$$

By reducing the data dimension, we reduced the number of parameters of the learning models so that we could build the models more efficiently.

4. Experiments

We targeted the problem of the stock prediction function approximation using the neural networks. We reduced the initial training data using the feature selection technique described in section 3, and presented the efficiency of the neural network learning using the reduced training data. We also gave the effectiveness of the stock prediction function trained with the reduced data.

4.1 Training Data

Initial training data consists of 137 feature values and one target value which are real values normalized between -1 and 1. We constructed the initial training data with 30,744 data from the two Korean stock markets, KOSPI and KOSDAQ. The data covers the period from January 1999 to December 2000. The data is organized as follows:

$$td_i = (f_i^1, f_i^2, \ldots, f_i^{137}, t_i, code, date),$$

where t_i is the target value, *code* is the code number of the stock, and *date* is the trading date. The features are the technical indicators assumed to affect the changes of the stock price. For example, the moving average of the price, the moving average of the trading volume and the trend transition are the candidate features. The target value is the rate of change of the next day. The positive target value means the increase of the price and the negative value expresses the decrease. There are about 2000 stocks in the Korean stock market, and the number of the data is too large for the training period. We built the initial training data for only the stocks that match the specific patterns in [3].

We reduced the initial training data through the **data filtering** step that uses the selected intervals of the features and the **dimension reduction** step that describes the data with the selected features only. Table 1 shows the changes of the training data size and the dimensionality of the data during the training data reduction. The initial training data was reduced to one third in size through the reduction steps with the feature selection.

Table 1. Reduction of the initial training data

	# of data	Dimension	Size (KB)
Initial Data	30,744	140	50,500
Data Filtering	30,430	140	49,925
Dimension Reduction	30,430	45	17,266

4.2 Efficiency of the Learning

The reduction of the training data with the feature selection makes the training fast. The neural network learning consists of two steps. In the first step, the network is trained with 1000 epochs and we determine the number of epochs which provides minimum training error. We call this the **first learning**. At the next step, the function of the stock price prediction is generated through the training with the above epoch, which is called the **second learning**. Table 2 shows the neural network training times.

Table 2. Change of the neural network training times (min)

	1st Learning	2nd Learning	Average of the 2nd
Initial Data	460	150	0.52
Reduce Data	270	32	0.25

The numbers of epochs for the second learning are 289 for the initial training data and 127 for the reduced data set. Therefore, the figures in the 3rd column don't represent the reduction of the training time. We present the average training time per epoch in the 4th column in Table 2. Table 2 suggests that the proposed feature selection technique contributes to the decrease of the training time.

4.3 Effectiveness of the Trained Models

We compared the performance of the functions generated with the initial training data and the reduced data set. We calculated the average profits per trade (PPT), the accuracy, and the number of trades for the period between January 2, 2003 and November 30, 2004. The accuracy is the ratio of the trades with profit to the total trades.

The simulation of the stock trading is performed on both KOSPI and KOSDAQ, and it uses a specific trading policy [4]. In the simulation, we invested 10,000,000 Won per trade. The PPT, accuracy and the number of trades depend on the trading policy. Table 3 and 4 show the simulation results that meet some criteria[§]. The results were selected whose PPT was over 2.7%[**], accuracy was more than 55%, and the number of trades was over 27, which means that there are 1.5 trades a month on average. In the following, the stock price prediction *function1* is estimated using the initial training data and the *function2* is approximated with the reduced data set.

Table 3. Simulation results with *function1*

Results	PPT(%)	Accuracy(%)	#	Profit(W)
1	3.18	58.1	31	8,533,988
2	3.12	57.1	28	7,222,659
3	2.91	64.3	28	6,761,850
4	3.26	55.6	27	7,897,911
5	2.9	70.4	27	6,299,393

Table 4. Simulation results with *function2*

Results	PPT(%)	Accuracy(%)	#	Profit(W)
1	4.15	69.7	33	12,235,590
2	3.65	81.8	33	10,020,770
3	4.4	66.7	28	10,394,221
4	3.16	57.1	49	13,332,804

For the performance comparison, Table 5 shows the averages and the degree of the performance improvement (*PE*), which is calculated as follows:

$$PE = \frac{new\_value - old\_value}{new\_value} \times 100\,(\%).$$

Table 5. Average performance

	Function1	*Function2*	*PE*
PPT	3.11	3.84	23.5%
Accuracy	61.1	68.9	12.8%
#	28.2	35.5	25.9%
Profit	7,343,160	11,495,846	56.6%

From Table 5, we can know that the performance improves by 23.5%, 12.8%, 25.9%, and 56.6%, in terms of the PPT, the accuracy, the number of trades, and the profit. The results support that the proposed feature selection technique well eliminates irrelevant features to the stock price prediction and contributes to better approximation of the prediction function that results in performance improvements.

5. Conclusion

This paper proposes the feature selection technique based on distributional differences for improving the efficiency and the effectiveness of the machine learning. This work was motivated by the idea that the effective data for the classification problem would have different distributional properties in each positive and negative data set. The reduction of the training data by the proposed technique helps to speed up the training and estimate the parameters of the learning models with more accuracy,

[§] 'W' means Korean Won.

[**] The PPT excludes the transaction cost (0.5% per trade).

which improves the performance of the trained models. The proposed feature selection technique is a kind of the filter model and the supervised method, and can select relevant features more quickly.

In the experiments, the proposed technique decreases the training time by about 40% by reducing the training data by about 30% in size. In the simulated stock trading, the technique also improves the performance of the trained function by about 24% in PPT, 13% in accuracy, and 56% in profit. The experimental results show that the proposed technique could contribute to improving the efficiency of the learning and the effectiveness of the trained models.

In order to discretize the real-valued features, we divided the range of the values into the 10 intervals with the same size. The size and the number of intervals may affect the efficiency and the effectiveness of the learning, so a study of the appropriate size and number of intervals will be required.

References

[1] H. Liu and H. Motoda, "Feature Selection for Knowledge Discovery and Data Mining", Kluwer Academic Publishers, 1998.

[2] Daphne Koller and Mehran Sahami, "Toward Optimal Feature Selection", In Proceedings of the 13th ICML, pp. 284-292, 1996.

[3] Sung-Dong Kim, Jae Won Lee, Jongwoo Lee, and Jinseok Chae, "A Two-Phase Stock Trading System Using Distributional Differences", In Proceedings of the 13th DEXA, LNCS 2453, pp. 143-152, 2002.

[4] Sung-Dong Kim, Jae Won Lee, "Induction of Stock Trading Rules Using Distributional Differences", In Proceedings of Korea Data Mining Conference, pp. 206-216, 2001.

[5] N. Wyse, R. Dubes, and A.K. Jain, "A critical evaluation of intrinsic dimensionality algorithms", In E.S. Gelsema and L.N. Kanal, editors, *Pattern Recognition in Practice*, Morgan Kaufmann Publishers, Inc., pp. 415-425, 1980.

[6] A.L. Blum and P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, pp.245-271, 1997.

[7] J.G. Dy and C.E. Brodley, "Feature subset selection and order identification for unsupervised learning", In Proceedings of the 17th International Conference on Machine Learning, pp. 247-254, 2000.

[8] Yiming Yang and Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", In Proceedings of the 14th International Conference on Machine Learning, pp. 412-420, 1997.

[9] G.H. John, R. Kohavi, and K. Pfleger, "Irrelevant feature and the subset selection problem", In Proceedings of the 11th International Conference on Machine Learning, pp. 121-129, 1994.

[10] M. Dash and H. Lie, "Feature selection methods for classification", Intelligent Data Analysis, Vol. 1, No. 3, pp. 131-156, 1997.

[11] L. Talavera, "Feature selection as a preprocessing step for hierarchical clustering", In Proceedings of International Conference on Machine Learning, pp. 389-397, 1999.

[12] K. Kira and L. Rendell, "A practical approach to feature selection", In Proceedings of the 9th ICML, pp 249-256, 1992.

[13] J.G. Dy and C.E. Brodley, "Feature subset selection and order identification for unsupervised learning", In Proceedings of the 17th International Conference on Machine Learning, pp. 247-254, 2000.

[14] H. Almuallim and T.G. Dietterich, "Learning with many irrelevant features", In Proceedings of the 9th National Conference on Artificial Intelligence, pp. 547-552, 1991.

Sung-Dong Kim

He received a Ph.D. degree in Computer Engineering from Seoul National Univ. in 1999. He has been an assistant professor at Hansung Univ. since 2001. He played an active role in developing the commercial English-Korean machine translation systems, such as Enkor, E-Tran, and SmarTran. His research interests are in the area of Machine Translation, Natural Language Processing, Data Mining, Computational Finance, and Machine Learning.