Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 327142, 9 pages
http://dx.doi.org/10.1155/2014/327142

Research Article

An Incremental Classification Algorithm for Mining Data with Feature Space Heterogeneity

Yu Wang[1,2]

[1] *School of Economic and Business Administration, Chongqing University, Chongqing 400030, China*
[2] *Chongqing Key Laboratory of Logistics, Chongqing University, Chongqing 400044, China*

Correspondence should be addressed to Yu Wang; yuwang@cqu.edu.cn

Received 16 December 2013; Accepted 13 January 2014; Published 19 February 2014

Academic Editor: J. J. Judice

Feature space heterogeneity often exists in many real world data sets so that some features are of different importance for classification over different subsets. Moreover, the pattern of feature space heterogeneity might dynamically change over time as more and more data are accumulated. In this paper, we develop an incremental classification algorithm, Supervised Clustering for Classification with Feature Space Heterogeneity (SCCFSH), to address this problem. In our approach, supervised clustering is implemented to obtain a number of clusters such that samples in each cluster are from the same class. After the removal of outliers, relevance of features in each cluster is calculated based on their variations in this cluster. The feature relevance is incorporated into distance calculation for classification. The main advantage of SCCFSH lies in the fact that it is capable of solving a classification problem with feature space heterogeneity in an incremental way, which is favorable for online classification tasks with continuously changing data. Experimental results on a series of data sets and application to a database marketing problem show the efficiency and effectiveness of the proposed approach.

1. Introduction

In classification problems, feature space heterogeneity is the phenomenon that a data set consists of some heterogeneous subsets, and the optimal features for classification are distinct over different subsets. The challenge of this problem is that we do not know how many heterogeneous subsets exist in the data set or which subset each sample belongs to. In the last decade, the problem of feature space heterogeneity in data has been addressed under different names, such as local feature relevance [1], case-specific feature weights [2], relevance in context [3], feature space and class heterogeneity [4], and attribute instability [5].

Feature space heterogeneity exists widely in various application fields of classification techniques, such as marketing, customs inspection decision, credit scoring, and medical diagnosis. For example, in marketing, a major concern of the market managers is to develop and implement efficient marketing programs by fully utilizing the customer databases and identifying the households that are most likely to be interested in the marketing programs. The above process can be formulated as a classification problem, in which the features (attributes) are characteristics of the households such as demographic, psychographic, and behavioral information, and the target variable is whether a household responds to the marketing messages. After the responding probability of each household is predicted by using certain classification techniques, the marketing messages are sent to those households with the highest probabilities. However, it is stated in Allenby and Rossi [6] that, as consumer preferences and sensitivities become more diverse, it becomes less and less efficient to consider the market in the aggregate. Desarbo et al. [7] also argue that predictions typically made with a single set of parameter values may not fully capture individual consumer differences in the sample. In other words, relevant features for predicting the responding probabilities may vary in different groups of customers. Therefore, it is important to take the feature space heterogeneity into consideration when solving the above marketing problems. Other similar examples can be found in customs inspection decision [8], medical diagnosis [9], rushes editing [10], and accident analysis [11].

If significant feature space heterogeneity exists in the data set and global feature selection is implemented for constructing a classification system, the resulting model is inevitably diffused by an averaging effect over the entire problem, since the best features on which to base the classification model vary in different subsets [4]. Therefore, dealing with feature space heterogeneity is an important issue in classification problems. Apte et al. [4] suggest that a logical first step is to decompose the classification problem with feature space heterogeneity into its constituent subproblems. In their study, an Important Profile Angle (IPA) is defined to indicate the degree to which the importance of each feature varies between two subproblems. The IPA is then used to guide the data set partitioning in an iterative way. However, the practical stopping criterion is still under investigation, and these ideas are not readily applicable for classification problems with numeric features. Therefore, an effective approach to classification problems with feature space heterogeneity is needed.

On the other hand, in some real world applications of classification techniques, new data are presented in sequence and added to the historical data set. Consequently, feature space heterogeneity existing in the historical data set might dynamically change over time. For example, in the custom inspection decision problem, there are thousands of declared goods waiting to be exported or imported every day. The custom officials have to decide whether an inspection is needed for declared goods by solving a classification problem [8]. Due to the variety and diversity of export/import trades, even in the same merchandise category, the relevant features may vary in different subcategories, as we have found in a research project sponsored by China Customs. Therefore, feature space heterogeneity is exhibited in this classification problem. Meanwhile, as more and more historical data are accumulated in the data base of customs, the underlying feature space heterogeneity might change and, accordingly, the classifier has to be updated for better accuracy. However, if we reconstruct the classifier once a batch of new data comes into the data base, the computational burden would be high due to the continuity of data stream, and the information and patterns learned in the past would be wasted. Moreover, constraints on the time and resource for processing the data could hardly be met. To deal with this problem, the classification approach should be capable of incremental learning, which is an active research direction [12]. The main characteristics of incremental learning are [13] as follows: (1) examples are not all available a priori but become available over time, usually one at a time; (2) since learning may need to go on (almost) indefinitely, a classifier needs to respond quickly in an online manner and process the data in a continuous way.

In this paper, we develop a novel classification approach, Supervised Clustering for Classification with Feature Space Heterogeneity (SCCFSH), to address the above problems, that is, feature space heterogeneity and incremental learning. Our approach is based on the ECCAS algorithm proposed by Li and Ye [14]. The main idea of our approach is to first divide the sample set into a number of subsets by supervised clustering such that samples in each subset are with the same class label and then calculate the relevance of features in each subset. The feature relevance is then incorporated in calculating the distances used for classification. The main advantage of SCCFSH lies in the fact that it is capable of solving a classification problem with feature space heterogeneity in an incremental way, which is favorable for online classification tasks with continuously changing data set. Experimental results on a series of data sets show that the proposed SCCFSH could achieve favorable classification performance and be capable of fast and incremental learning.

The rest of this paper is organized as follows. In Section 2, we briefly review the previous researches on feature space heterogeneity and incremental learning. Section 3 presents the classification approach SCCFSH we develop. The experimental results on a series of benchmark data sets and a real world application are reported in Section 4. Conclusion and discussion are made in Section 5.

2. Related Works

The phenomena that relevant features for classification vary across the data set have been observed by many researchers and practitioners [4, 15–17]. Until recently, a number of classification methods have been developed, which can be divided into two categories. In the first category, one of the best known methods is "bagging" [18]. In this approach, n subsets are generated by randomly sampling from the original set of samples. Consequently, relevant features might be different in the obtained subsets. Based on this approach, Puuronen et al. [19] proposed a Meta-Level Classification (MLC) method, which can be used to deal with the problem of feature space heterogeneity. MLC first divides the training sample set into some subsets and obtains the component classifiers based on these subsets. In the application phase, testing samples are put into the training sample set, and MLC dynamically selects the optimal component classifier for a testing sample by comparing the performance of different classifiers in its neighborhood. Different from the method of sample partitioning, the Random Subspace Method (RSM) [20] divides the whole feature set into a number of feature subsets and constructs different classifiers based on the whole training samples with different feature subsets obtained. Feature space heterogeneity in testing samples is considered through synthesizing (usually by voting) the application results of all classifiers. These methods deal with the problem of feature space heterogeneity by firstly dividing sample set or feature set into different subsets in a random way and then training component classifiers in the subsets. These component classifiers are then combined for classification, mostly by major voting or selecting the optimal one. A major problem of these methods lies in the random set (sample set or feature set) partitioning, which may result in seriously biased component classifiers due to the feature redundancy and irrelevance in some subsets, especially for high dimensional data sets.

In the second category, modified lazy learning methods are applied to classification problems with feature space

heterogeneity. Friedman [1] addresses the problem of feature space heterogeneity by investigating the variability of feature relevance in different data subsets. In his method, the local relevance of features in each subset is measured by the estimated reduction in classification error. Hastie and Tibshirani [15] develop an adaptive form of nearest neighbor classification method for dealing with feature space heterogeneity. In their approach, distance metric for each sample is adaptively calculated in an iterative process using local discriminative information of features. Therefore, different relevant features are taken into account for classification in different subsets. Although both works report favorable results on their local approaches compared to global ones, both of them are computationally expensive [16]. Paredes and Vidal [21] propose a locally weighted lazy learning approach for better classification accuracy. In their method, different samples would have different feature weights obtained by approximately minimizing the Leaving-One-Out (LOO) classification error of the given training set. However, the computational complexity of this method is high because of the gradient descent algorithm employed to search for the optimal weights.

In spite of the fact that many researches have been carried out for dealing with feature space heterogeneity in classification, we have not found any for incremental learning among them. Researches on incremental classification are mainly focused on statistical methods [22], neural networks [23–25], and evolutionary algorithm [26]. Instance-based learning, especially k-nearest neighbor (k-NN) learning, is a widely used nonparametric incremental classification approach where training or learning does not take place until a query is made. In contrast to complex learning algorithms such as neural networks or support vector machines, k-NN learning does not require a complex function fitting process or model training procedure. Thus, it is easy to do incremental learning [27]. Nevertheless, once a query point with unknown class label is presented, conventional k-NN learning traverses the whole data set to find the k nearest neighbors of the query point. Therefore, the computational time and requirement of computer storage space of k-NN are not scalable to large amounts of data. To solve this problem, Li and Ye [14] propose a data mining algorithm based on supervised clustering to learn data patterns and use these patterns for classification. This algorithm enables a scalable and incremental learning of patterns from data with both numeric and nominal variables. However, it calculates the feature relevance by using squared correlation coefficient between predictor variables and target variable over the entire data set, regardless of the possible heterogeneity that exists in the feature space.

3. The Proposed Approach

The Supervised Clustering for Classification with Feature Space Heterogeneity (SCCFSH) proposed in this paper is based on the ECCAS [14]. However, SCCFSH differs significantly from ECCAS in that it takes feature space heterogeneity into consideration. SCCFSH first divides the data set into a number of subsets in a supervised way and then explores the feature relevance in each subset obtained. The main steps of SCCFSH include grid-based supervised clustering, supervised grouping of clusters, removal of outliers, calculation of feature relevance in each cluster, and distance-based classification.

3.1. Grid-Based Supervised Clustering. Consider a data set $S = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ in which $\mathbf{X}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ $(i = 1, 2, \ldots, N)$ is a p-dimensional sample. The label set of samples in DS is $L = \{y_1, y_2, \ldots, y_N\}$ $(y_i \in \{0, 1, \ldots, C\})$. Without loss of generality, we only consider the binary classification problem; that is, $y_i \in \{0, 1\}$, $(i = 1, 2, \ldots, N)$.

The grid-based supervised clustering procedure first divides the p-dimensional space of samples into grid cells and then generates clusters within the grid cells, as suggested in Li and Ye [14]. This procedure aims to avoid the problem that different presentation order of the same data points may generate different cluster structures. For example, a number of data points of the same class may appear consecutively in the data set. Without the above grid-based procedure, these data points would be grouped into one cluster, even though they are not close to each other at all. Consequently, the cluster structure is not robust to the presentation order of data points. With the above grid-based procedure, these data points can be prevented from joining into one single cluster.

In the proposed SCCFSH, the procedure of grid-based supervised clustering is similar to that in ECCAS. The main difference is that SCCFSH employs grid-based supervised clustering only for decomposing the classification problem into its constituent subproblems, and feature relevance in individual subproblems is considered in later steps. Thus, we simply use the conventional Euclidean distance metric

$$d\left(\mathbf{X}_i, \mathbf{X}_j\right) = \left(\sum_{d=1}^{p} \left|x_{i,d} - x_{j,d}\right|^2\right)^{1/2} \tag{1}$$

in supervised clustering without imposing any weights on the features in distance calculation. In comparison, ECCAS first calculates feature relevance by using the squared correlation coefficient $r_{d,y}^2$ between predictor variable x_d $(d = 1, 2, \ldots, p)$ and class label y over the entire data set, and then it incorporates them in distance calculation by using the following weighted Euclidean distance metric:

$$d\left(\mathbf{X}_i, \mathbf{X}_j\right) = \left(\sum_{d=1}^{p} r_{d,y}^2 \left(x_{i,d} - x_{j,d}\right)^2\right)^{1/2}. \tag{2}$$

In other words, ECCAS does not take any possible feature space heterogeneity into consideration.

3.2. Supervised Grouping of Clusters and Removal of Outliers. Supervised grouping of clusters plays an important role in the proposed SCCFSH. If some underlying clusters cover the area of several grid cells, grid-based supervised clustering would divide these large clusters into small clusters. Therefore, refinement of the clustering results is needed. In

INPUT: Clusters $C_1, C_2, \ldots, C_k$ obtained by grid-based supervised clustering.
OUTPUT: $\mathfrak{R} = \{C_1^{\text{New}}, C_2^{\text{New}}, \ldots, C_m^{\text{New}}\}$.
(1) Set $\mathfrak{R} = \varnothing$, $\mathfrak{R}_0 = \{C_1, C_2, \ldots, C_3\}$ and $t = 1$;
(2) Among all possible pairs of clusters (C_r, C_s) in $\mathfrak{R}_{t-1}$ find the one, say (C_i, C_j), such that $d(C_i, C_j) = \min_{r,s} \{d(C_r, C_s)\}$;
(3) **IF** Label (C_i) = Label (C_j) **THEN**
set $C_q = C_i \bigcup C_j$ and produce the new clustering $\mathfrak{R} = (\mathfrak{R}_{t-1} - \{C_i, C_j\}) \bigcup C_q$;
ELSE $\mathfrak{R}_t = \mathfrak{R}_{t-1} - \{C_i, C_j\}$, $\mathfrak{R} = \mathfrak{R} \bigcup C_i \bigcup C_j$;
Set $t = t + 1$;
(4) Repeat steps 2-3 until $\mathfrak{R}_t = \mathfrak{R}_{t-1}$, return $\mathfrak{R}$.

Algorithm 1: Main procedure of supervised grouping of clusters.

SCCFSH, we iteratively group the clusters obtained from supervised clustering in the same way as ECCAS. In this grouping procedure, a single linkage method is used, where the distance between two clusters is defined as the distance between their nearest "points" [28]:

$$d(C_q, C_s) = \min \{d(C_i, C_s), d(C_j, C_s)\}. \quad (3)$$

In (3), C_q is the newly formed cluster by merging clusters C_i and C_j and C_s is an old cluster. Note that, in (3), the Euclidean distance metric represented in (1) is still used for distance calculation in SCCFSH, since our main aim is to decompose the classification problem into its subproblems according to the spatial distribution of samples. The main procedure of supervised grouping of clusters is shown in Algorithm 1. In step 3 of Algorithm 1, Label (C_i) refers to the class label of cluster C_i.

By applying the grid-based supervised clustering and supervised grouping of clusters, a number of data groups can be obtained. In each group, samples are with the same class label. Nevertheless, in some groups, there may be only a few samples. These groups often represent noises (outliers) in training data samples and thus should be removed. A common way is to check the number of samples in a cluster. Those clusters whose number of samples is no more than the threshold are removed from the clusters. However, how to choose this threshold strongly depends on the characteristic of the data set. For example, in some data set where the number of samples in one class is much larger than that in another class, the threshold value should be different for clusters with different class labels. In this paper, we simplify our work by setting this threshold value to be equal to 1.

3.3. Calculation of Feature Relevance for Classification. After the above three steps, we could obtain a number of clusters, and samples in each cluster are with the same class label. Vucetic and Obradovic [29] argue that some features with the same values may result in quite different outputs (class labels) in different regions. Therefore, spatial characteristics of samples should be explored for better classification performance. Motivated by this statement, we investigate the feature relevance in the obtained clusters that represent different spatial distributions of samples. It is stated in Gennari et al. [30] that features are relevant if their values vary systematically with class membership. Theodoridis and Koutroumbas [28] also argue that relevant features would have large between-class variance and small within-class variance. Accordingly, in our approach, the variances of features in a cluster are utilized to calculate their relevance to the class label. If the variance of a feature is small, it is implied that this feature is informative for the class label and thus should carry more weight in classification. Therefore, we calculate the relevance of feature x_d in the cluster C_k by

$$r_k(d) = \frac{1}{(1 + \text{Var}(x_d))}. \quad (4)$$

Once the relevance of each feature $r_k(x_d)$ $(d = 1, 2, \ldots, p)$ in C_k is obtained, we standardize them by requesting that the sum of $r_k(x_d)$ with respect to d be equal to p, in accordance with the sum of weights in conventional Euclidean distance metric (the weight of each feature is 1). The relevance of feature x_d in cluster C_k is finally determined by

$$R_k(d) = \frac{p \cdot r_k(d)}{\sum_{d=1}^{p} r_k(d)}. \quad (5)$$

With all the cluster centroids $\widehat{\mathbf{X}}_1, \widehat{\mathbf{X}}_2, \ldots, \widehat{\mathbf{X}}_k$ and feature relevance $R_k(x_d)$ $(d = 1, 2, \ldots, p)$ in each cluster, we could consider each cluster centroid as a sample in the feature space and classify a new sample $\mathbf{X}$ simply using the nearest neighbor (NN) rule by defining the weighted distance metric as

$$d(\mathbf{X}, \widehat{\mathbf{X}}_k) = \left(\sum_{d=1}^{p} R_k(d) \left| x_d - \widehat{x}_{k,d} \right|^2 \right)^{1/2}, \quad (6)$$

where $R_k(d)$ are defined in (4). Clearly, if the relevance of a feature is strong, it carries high weight in distance calculation for classification.

In summary, the main steps of the proposed SCCFSH are shown in Algorithm 2.

4. Experimental Study

To verify the effectiveness of the proposed approach for classification problems with feature space heterogeneity, we

Table 1: Benchmark data sets used in the experiments.

Benchmark data set	Number of classes	Number of features	Number of samples	Area
German	2	20	700	Financial
Waveform	2	21	400	Physical
Ringnorm	2	20	400	N/A
Twonorm	2	20	400	N/A
Image	2	18	1300	N/A
Statlog (Shuttle)	6	9	58000	Physical
Magic gamma	2	10	19020	Physical
Yeast	10	8	1484	Life
Ecoli	2	7	336	Life
Abalone	3	8	4177	Life
Page-Blocks	5	10	5473	Computer
Letter Recognition	26	16	20000	Computer
Gauss_8D	2	8	5000	ELENA

INPUT:
Training data set $S_1 = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ with label set $L = \{y_1, y_2, \ldots, y_N\}$;
Testing data set $S_2 = \{\mathbf{X}_{N+1}, \mathbf{X}_{N+2}, \ldots, \mathbf{X}_{N+M}\}$ with unknown labels.
OUTPUT:
Predicted labels $\widehat{y}_{N+1}, \widehat{y}_{N+2}, \ldots, \widehat{y}_{N+M}$ for S_2.
(1) Apply the grid-based supervised clustering to S_1 and L;
(2) Apply the algorithm shown in Algorithm 1 to Clusters $C_1, C_2, \ldots, C_k$ obtained in step 1;
(3) Calculate the relevance of each feature $R_k(d), (d = 1, 2, \ldots, p)$
in $C_1^{\text{New}}, C_2^{\text{New}}, \ldots, C_m^{\text{New}}$ obtained in step 2;
(4) **LOOP** for each data point $\mathbf{X}_j, (j = N+1, N+2, \ldots, N+M)$ in S_2:
(i). Calculate the distance between $\mathbf{X}_j$, and C_i^{New}, $(i = 1, 2, \ldots, m)$ obtained in step 2,
by using the distance metric defined in (6);
(ii). Set $\widehat{y}_j$ as the label of the nearest cluster;
(5) Output the predicted labels $\widehat{y}_{N+1}, \widehat{y}_{N+2}, \ldots, \widehat{y}_{N+M}$ for S_2.

Algorithm 2: Main steps of SCCFSH.

implement two sets of experiments. In the first set of experiments, we artificially construct some mixed data sets with feature space heterogeneity and show that SCCFSH is effective and time-efficient. In the second set of experiments, we apply the proposed classification algorithm to a real world customer targeting problem.

4.1. Experiments on Benchmark Data Sets. In this set of experiments, we select 13 benchmark data sets from the ELENA project [31] and UCI machine learning repositories (available at http://archive.ics.uci.edu/ml/ and https://www.elen.ucl.ac.be/neural-nets/Research/Projects/ELENA/elena.htm). Basic characteristics of these selected data sets are briefly summarized in Table 1.

As shown in Table 1, these benchmark data sets are from different backgrounds and have different number of features with different units of feature values. To avoid the dominance of features with large values over those with small values, we standardize the selected data sets such that the value of each feature satisfies a random distribution with zero mean and unit standard deviation. In this paper, we mainly focus on binary classification problems. Thus, for data sets with more than two classes, we convert them into a number of two-class subproblems and choose one of the resulting subproblems for the experimental study. For example, the data set *Letter Recognition* has 26 classes. We select a subset of samples in "A" and "D" classes and denote it as *Letter Recognition* (A versus D).

Since the main aim of this study is to investigate the effectiveness of our classification algorithm on data sets with significant feature space heterogeneity, we construct 12 mixed data sets by merging some benchmark data sets selected from the data sets listed in Table 1. In case the component (benchmark) data sets to be merged have different numbers of features, we add some random features into the data sets with fewer features to achieve equal dimensionalities. Values of each added random feature are generated from a normal distribution with zero mean and unit standard deviation. Because some data sets have much more samples than others, we randomly select 1,000 samples from those component data sets with more than 1,000 samples to balance the component proportions in each mixed data set. Structures of the 12 mixed data sets are described in Table 2.

Table 2: Structures of the mixed data sets.

Mixed data set	Component data sets
Data set 1	Abalone and Gauss_8D
Data set 2	Image and German
Data set 3	Letter and Image
Data set 4	Magic and Page-blocks
Data set 5	Shuttle and Magic
Data set 6	Yeast and Ecoli
Data set 7	German and Ringnorm and Twonorm
Data set 8	Twonorm and Waveform and Ringnorm
Data set 9	German and Image and Waveform
Data set 10	Magic and Gauss_8D and Page-blocks
Data set 11	Yeast and Abalone and Gauss_8D
Data set 12	Magic and Shuttle and Abalone

To investigate the possible heterogeneity that exists in the feature space of these mixed data sets, we evaluate the feature relevance in each component data set using the squared correlation coefficients between the features and class label. The results are shown in Table 3.

Table 3 indicates that the order of feature relevance varies in different component (benchmark) data sets. Therefore, it can be concluded that most mixed date sets will have significant feature space heterogeneity, which is suitable for our experimental study.

We next apply the proposed SCCFSH to the mixed data sets. To demonstrate the effectiveness of SCCFSH on classification problems with feature space heterogeneity, we compare its performance with that of the ECCAS and Class Prototype Weight (CPW) learning method proposed by Paredes and Vidal [21]. In CPW learning, different training samples have different optimal feature weights. These weights are determined by approximately minimizing the Leaving-One-Out NN classification error of the given training set. Therefore, CPW learning can be employed for classification problems with feature space heterogeneity. In CPW learning, there are several parameters that need to be set. As suggested in Paredes and Vidal [21], we set $\beta = 8$, $\mu_{ij} = 0.005$, and $\rho_i = 0.005$.

In the experiments, B-Fold Cross Validation [32] is applied to estimate the error rates. Each mixed data set is first divided into B subsets randomly, and $B-1$ subsets are used as the training set and the remaining subset is used as the testing set. For simplicity, we arbitrarily set the $B = 5$. Training-testing experiment for each mixed data set is run 30 times using different random 5-fold partitions. The classification error rates of ECCAS, CPW, and SCCFSH averaged over 30 runs and the results of paired t-test are shown in Table 4.

It can be observed from Table 4 that, in comparison with ECCAS and CPW learning method, the proposed classification approach SCCFSH could obtain comparative or equivalent results in 12 mixed data sets. It is noteworthy that, in comparison to ECCAS, the proposed SCCFSH achieves uniformly better classification performance over most (11 out of 12) data sets. This may be attributed to the fact that SCCFSH takes the feature space heterogeneity that exists in the mixed data sets into consideration.

Since one main aim of our study is to develop a classification approach with incremental learning ability, we compare the average computational times of SCCFSH and CPW on each mixed data set over 30 runs. The results are shown in Table 5.

Table 5 shows that the average computational time of our SCCFSH is much less than that of the CPW learning. This is because CPW learning employs gradient descent method to search for approximately optimal feature weights by minimizing the LOO NN error rate. After these weights are adjusted in one iteration, the algorithm has to traverse the whole data set to find the nearest neighbors for each sample in the next iteration. Besides, the convergence rate for the gradient descent method is dependent on the condition number of the Hessian (the ratio of the largest eigenvalue to the smallest one of the Hessian) and can be very slow, as presented, for example, by Luenberger and Ye [33]. In contrast, the computational complexity of SCCFSH is $O(pNM)$, the same as that of ECCAS, where p is the number of features, N is the number of samples, and M is the number of clusters.

4.2. Application to a Real World Customer Targeting Problem. The data set used in this set of experiments is taken from a solicitation of 9822 European households to buy insurance for their recreational vehicles (RV) (available online at http://www.liacs.nl/~putten/library/cc2000/). In this data set, each household's record contains a target variable indicating whether they buy insurance and 93 predictor variables indicating information on both sociodemographic characteristics and ownership of various types of insurance policies. A more detailed description of the data set is presented in Kim and Street [34].

In the experiments, we use two separate data sets: a training set with 5,822 households and an evaluation set with 4,000 households. Of the 5,822 prospects in the training data set, 348 purchased RV insurance, resulting in a hit rate of $348/5822 = 5.97\%$. From the manager's perspective, he/she would like to increase this hit ratio by selecting those households with highest responding probabilities and sending mails to them. Therefore, efficient classification model based on the training set is needed to predict the responding probability of each household. The evaluation data is used to validate the predictive classification model. Hereafter, we define the households who purchase and do not purchase the RV insurance as positive and negative, respectively. Therefore, through supervised clustering, the households are partitioned into a number of clusters which are either positive or negative.

Since we are interested in the top i% of customers with the highest probability to buy RV insurance in the evaluation data set, the method's predictive accuracy is examined by computing the hit rate among the selected households. Consequently, we modified the SCCFSH shown in Algorithm 2 as follows. Instead of finding the nearest neighbor of each sample $\mathbf{X}_j$ ($j = N+1, N+2, \ldots, N+M$) in the testing data set S_2, we find two nearest neighbors (clusters) from different classes. Denote by

Table 3: Feature relevance in component data sets.

Component data set	Order of feature relevance (from strongest to weakest)
German	18, 10, 17, 15, 11, 4, 19, 8, 20, 9, 16, 14, 13, 7, 12, 5, 6, 3, 2, 1
Waveform	1, 21, 2, 12, 3, 13, 11, 4, 20, 5, 6, 19, 10, 7, 14, 17, 18, 16, 9, 8, 15
Ringnorm	11, 16, 14, 12, 4, 19, 15, 7, 6, 10, 5, 3, 9, 20, 18, 2, 1, 17, 13, 8
Twonorm	12, 10, 7, 3, 18, 13, 19, 20, 2, 8, 6, 16, 14, 5, 17, 9, 1, 15, 4, 11
Image	5, 6, 17, 4, 3, 8, 7, 13, 2, 1, 12, 16, 9, 10, 11, 14, 15, 18
Statlog (Shuttle)	4, 6, 2, 3, 5, 7, 8, 9, 1
Magic gamma	4, 5, 8, 10, 3, 6, 7, 2, 1, 9
Yeast	4, 5, 6, 7, 3, 2, 1, 8
Ecoli	4, 3, 2, 5, 1, 7, 6
Abalone	5, 4, 8, 6, 7, 3, 1, 2
Page-Block	8, 3, 6, 9, 2, 7, 10, 1, 5, 4
Letter Recognition	3, 4, 2, 16, 10, 1, 13, 5, 15, 6, 14, 8, 11, 7, 12, 9
Gauss_8D	1, 6, 5, 2, 4, 8, 7, 3

Table 4: Average error rates of ECCAS, CPW, and SCCFSH over 30 runs.

Mixed data set	Average error rate			Sig. (paired t-test)[a]
	ECCAS	CPW	SCCFSH	
1	43.65% ± 4.06%	44.72% ± 5.31%	41.63% ± 4.46%	0.045
2	13.14% ± 2.85%	9.39% ± 2.42%	11.20% ± 3.36%	0.008
3	10.61% ± 0.98%	9.44% ± 2.16%	10.14% ± 1.35%	0.081
4	26.21% ± 3.94%	25.52% ± 4.65%	25.00% ± 7.11%	0.681
5	20.17% ± 1.94%	17.05% ± 2.82%	16.17% ± 1.94%	0.111
6	35.95% ± 2.35%	37.32% ± 2.74%	35.75% ± 3.38%	0.776
7	37.99% ± 2.71%	36.67% ± 2.64%	35.41% ± 2.49%	0.040
8	30.92% ± 1.39%	31.92% ± 1.67%	30.76% ± 1.08%	0.653
9	20.19% ± 2.17%	16.46% ± 2.81%	17.87% ± 3.03%	0.086
10	37.07% ± 3.53%	35.15% ± 2.78%	36.62% ± 3.34%	0.059
11	45.06% ± 2.49%	47.76% ± 3.76%	45.98% ± 3.02%	0.019
12	28.82% ± 2.16%	26.98% ± 4.07%	28.16% ± 2.60%	0.165

[a]The method (ECCAS or CPW) with lower error rate versus SCCFSH.

$d(\mathbf{X}_j, C_{\text{pos}}^{\text{New}})$ and $d(\mathbf{X}_j, C_{\text{neg}}^{\text{New}})$ the distances between $\mathbf{X}_j$ and its nearest positive cluster and negative cluster, respectively. The probability of $\mathbf{X}_j$ belonging to a specific class (e.g., positive class) can be calculated by

$$\Pr\left(y_j = 1\right) = \frac{d\left(\mathbf{X}_j, C_{\text{neg}}^{\text{New}}\right)}{\left(d\left(\mathbf{X}_j, C_{\text{pos}}^{\text{New}}\right) + d\left(\mathbf{X}_j, C_{\text{pos}}^{\text{New}}\right)\right)}, \quad (7)$$

where $y_j = 1$ indicates that household j would buy RV insurance. Equation (6) means that a further distance between a testing sample and its nearest negative cluster implies higher probability that this testing sample belongs to positive class. Equation (6) is employed to modify the output of SCCFSH in application to the customer targeting problem in this experiment.

Similar to the evaluation mechanism of prediction accuracy in Kim et al. [35], we estimate the probability of buying new insurance for each household in the evaluation data with SCCFSH. After sorting the households in descending order of the estimated probability $\Pr(y = 1)$, we compute the cumulative hit rate of a model over various target points i where $i = 10, 20, \ldots, 50$. A comparison of cumulative hit ratios obtained by SCCFSH, ECCAS, and the method proposed in Kim et al. [35] is shown in Figure 1. Note that, for ECCAS, formula (7) is also used to modify the outputs to estimate the probability $\Pr(y = 1)$.

It can be observed from Figure 1 that the proposed SCCFSH shows uniformly better performance at target point $i = 5, 10, 15, 20, 25, 30$. Considering that, in this application, the market managers are more interested in targeting fewer customers with higher hit ratio, the result obtained by our approach is quite favorable. Moreover, our method takes much less time (less than 10 minutes on a computer with 1.5 GHz CPU and 256 M RAM) than that in ELSE/ANN (more than ten hours).

5. Conclusion

Feature space heterogeneity is the phenomenon that the optimal features for classification are distinct in different subsets of samples, but prior knowledge about these underlying

Table 5: Average computational times of SCCFSH and CPW learning.

Mixed data set	1	2	3	4	5	6	7	8	9	10	11	12
CPW	24.42 s	952.66 s	9.10 s	18.36 s	25.84 s	3.70 s	752.81 s	1174.24 s	703.33 s	8.92 s	13.01 s	8.84 s
SCCFSH	1.09 s	5.07 s	1.59 s	1.22 s	3.11 s	0.25 s	7.11 s	5.33 s	8.81 s	0.80 s	0.84 s	0.96 s

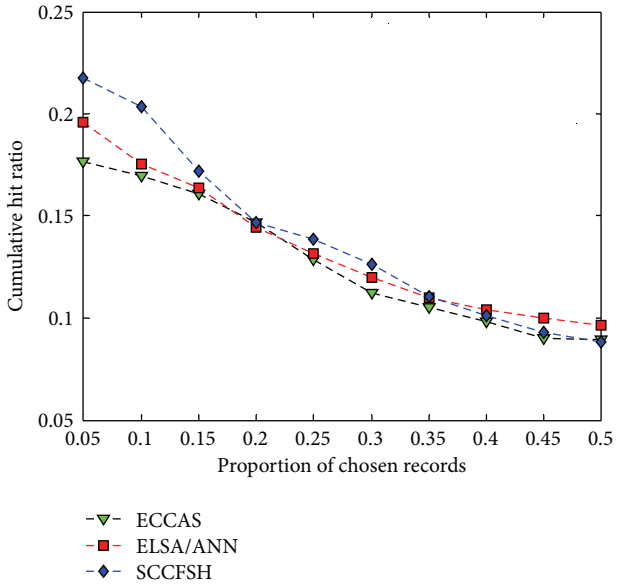

Figure 1: Cumulative hit ratios of ELSA/ANN, ECCAS, and SCCFSH.

subsets is unavailable. Moreover, in some real world applications of classification techniques, new data are presented in sequence and added to the historical data set after they are processed. Consequently, feature space heterogeneity existing in the historical data set might dynamically change and, accordingly, the classification system has to be updated for better accuracy. In this paper, we develop a Supervised Clustering for Classification with Feature Space Heterogeneity (SCCFSH) to address this problem. Our approach consists of four main steps: grid-based supervised clustering, supervised hierarchical grouping of clusters, feature relevance evaluation in each cluster, and weighted distance calculation for classification. The main advantage of the proposed SCCFSH is that it is enabled to deal with feature space heterogeneity in classification problems in a scalable and incremental way. Computational results in the experiments verify the efficiency and effectiveness of the proposed approach. In spite of the fact that we only consider binary classification problems in this paper, our approach can be easily extended to multiclass classification problems.

In the proposed SCCFSH, a cluster with only one sample is considered as an outlier and removed. When some samples from different classes overlap heavily in the data space, it might be inappropriate to consider a cluster with one data as an outlier. A possible direction for future research is to improve the proposed approach to dealing with overlapping samples.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The author is grateful to the editor and the anonymous reviewer for providing many helpful comments and suggestions, which have significantly improved the exposition and focus of this paper. This research is supported by the National Natural Science Foundation of China (NSFC Grant no. 71001112), the Fundamental Research Funds for the Central Universities (Project no. CQDXWL-2013-083), and the Social Science Research Fund for Young Teachers in Chongqing University (Project no. CDSK2009-11).

References

[1] J. H. Friedman, "Flexible metric nearest neighbor classification," Tech. Rep., Stanford University, 1994.

[2] C. Cardie and N. Howe, "Improving minority class prediction using case-specific feature weights," in *Proceedings of the 14th International Conference on Machine Learning*, 1997.

[3] P. Domingos, "Context-sensitive feature selection for lazy learners," *Artificial Intelligence Review*, vol. 11, no. 1–5, pp. 227–253, 1997.

[4] C. Apte, S. J. Hong, J. Hosking, J. Lepre, E. Pednault, and B. Rosen, "Decomposition of heterogeneous classification problems," *Intelligent Data Analysis*, vol. 2, no. 1, pp. 81–96, 1998.

[5] Z. Lazarevic, T. Fiez, and Z. Obradovic, "Adaptive boosting for spatial functions with unstable driving attributes," in *Knowledge Discovery and Data Mining. Current Issues and New Applications*, vol. 1805 of *Lecture Notes in Computer Science*, pp. 329–340, 2000.

[6] G. M. Allenby and P. E. Rossi, "Marketing models of consumer heterogeneity," *Journal of Econometrics*, vol. 89, no. 1-2, pp. 57–78, 1998.

[7] W. S. Desarbo, A. Ansari, P. Chintagunta et al., "Representing heterogeneity in consumer response models," *Marketing Letters*, vol. 8, no. 3, pp. 335–348, 1997.

[8] Z. Hua, S. Li, and Z. Tao, "A rule-based risk decision-making approach and its application in China's customs inspection decision," *Journal of the Operational Research Society*, vol. 57, no. 11, pp. 1313–1322, 2006.

[9] K. J. Cios and G. William Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, no. 1-2, pp. 1–24, 2002.

[10] Y. Liu, Y. Liu, and K. C. C. Chan, "Dimensionality reduction for heterogeneous dataset in rushes editing," *Pattern Recognition*, vol. 42, no. 2, pp. 229–242, 2009.

[11] J.-T. Wong and Y.-S. Chung, "Analyzing heterogeneous accident data from the perspective of accident occurrence," *Accident Analysis and Prevention*, vol. 40, no. 1, pp. 357–367, 2008.

[12] T. G. Dietterich, "Machine-learning research: four current directions," *AI Magazine*, vol. 18, no. 4, pp. 97–136, 1997.

[13] C. Giraud-Carrier, "A note on the utility of incremental learning," *AI Communications*, vol. 13, no. 4, pp. 215–223, 2000.

[14] X. Li and N. Ye, "A supervised clustering and classification algorithm for mining data with mixed variables," *IEEE Transactions on Systems, Man, and Cybernetics A*, vol. 36, no. 2, pp. 396–406, 2006.

[15] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 607–616, 1996.

[16] I. Scrypnyk and T. K. Ho, "Feature selection and training set sampling for ensemble learning on heterogeneous data," Tech. Rep., DIMACS, 2003.

[17] M. K. Lim and S. Y. Sohn, "Cluster-based dynamic scoring model," *Expert Systems with Applications*, vol. 32, no. 2, pp. 427–431, 2007.

[18] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[19] S. Puuronen, V. Terziyan, and A. Tsymbal, "A dynamic integration algorithm for an ensemble of classifiers," in *Foundations of Intelligent Systems*, vol. 1609 of *Lecture Notes in Computer Science*, pp. 592–600, 1999.

[20] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[21] R. Paredes and E. Vidal, "Learning weighted metrics to minimize nearest-neighbor classification error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1100–1110, 2006.

[22] S. M. Weiss and C. A. Kulikowski, *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*, Morgan Kaufmann, San Mateo, Calif, USA, 1991.

[23] K. Yamauchi, N. Yamaguchi, and N. Ishii, "Incremental learning methods with retrieving of interfered patterns," *IEEE Transactions on Neural Networks*, vol. 10, no. 6, pp. 1351–1365, 1999.

[24] S.-U. Guan and S. Li, "Incremental learning with respect to new incoming input attributes," *Neural Processing Letters*, vol. 14, no. 3, pp. 241–260, 2001.

[25] L. Su, S. U. Guan, and Y. C. Yeo, "Incremental self-growing neural networks with the changing environment," *Journal of Intelligent Systems*, vol. 11, no. 1, pp. 43–74, 2001.

[26] S.-U. Guan and F. Zhu, "An incremental approach to genetic-algorithms-based classification," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 35, no. 2, pp. 227–239, 2005.

[27] P. Kang and S. Cho, "Locally linear reconstruction for instance-based learning," *Pattern Recognition*, vol. 41, no. 11, pp. 3507–3518, 2008.

[28] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Elsevier, 2nd edition, 2003.

[29] S. Vucetic and Z. Obradovic, "Descovering homogeneous regions in spatial data through competition," in *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[30] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial Intelligence*, vol. 40, no. 1–3, pp. 11–61, 1989.

[31] C. Avilcs-Cruz, A. Guerin-Deguc, J. L. Voz, and D. Van Cappel, "Enhanced learning for evolutive neural architecture (ELENA)," Tech. Rep. R3-B1-P, Neural Network Group, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium, 1995.

[32] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[33] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, Springer, New York, NY, USA, 3rd edition, 2008.

[34] Y. Kim and W. N. Street, "An intelligent system for customer targeting: a data mining approach," *Decision Support Systems*, vol. 37, no. 2, pp. 215–228, 2004.

[35] Y. Kim, W. N. Street, G. J. Russell, and F. Menczer, "Customer targeting: a neural network approach guided by genetic algorithms," *Management Science*, vol. 51, no. 2, pp. 264–276, 2005.