Heat Pump Detection from Coarse Grained Smart Meter Data with Positive and Unlabeled Learning

Hongliang Fei, Younghun Kim, Sambit Sahu and Milind Naphade
IBM T.J. Watson Research, Yorktown Heights, NY 10598
{hfei,kimy,sambits,naphade}@us.ibm.com.com

Sanjay K. Mamidipallis
IBM Global Business Service
Atlanta, GA 30328
sanjay.mamidipalli@us.ibm.com

John Hutchinson
IBM Global Business Service
Piscataway, NJ 08854
John.Hutchinson@us.ibm.com

ABSTRACT

Recent advances in smart metering technology enable utility companies to have access to tremendous amount of smart meter data, from which the utility companies are eager to gain more insight about their customers. In this paper, we aim to detect electric heat pumps from coarse grained smart meter data for a heat pump marketing campaign. However, appliance detection is a challenging task, especially given a very low granularity and partial labeled even unlabeled data. Traditional methods install either a high granularity smart meter or sensors at every appliance, which is either too expensive or requires technical expertise. We propose a novel approach to detect heat pumps that utilizes low granularity smart meter data, prior sales data and weather data. In particular, motivated by the characteristics of heat pump consumption pattern, we extract novel features that are highly relevant to heat pump usage from smart meter data and weather data. Under the constraint that only a subset of heat pump users are available, we formalize the problem into a positive and unlabeled data classification and apply biased Support Vector Machine (BSVM) to our extracted features. Our empirical study on a real-world data set demonstrates the effectiveness of our method. Furthermore, our method has been deployed in a real-life setting where the partner electric company runs a targeted campaign for 292,496 customers. Based on the initial feedback, our detection algorithm can successfully detect substantial number of non-heat pump users who were identified heat pump users with the prior algorithm the company had used.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms, Experimentation

Keywords

Smart Meter Data Mining, Feature Extraction, Heat Pump Detection, Positive and Unlabeled Learning

1. INTRODUCTION

Recently smart metering infrastructure is being rapidly deployed in the United States and elsewhere. Smart meter data mining and related data management system have been popular in data management [13, 21, 26, 31] and data mining communities [14, 7, 28], e.g., energy disaggregation [14, 24], power theft detection [3, 20], etc. The smart meter data investigated therein is typically fine grained with a relatively small scale.

Unlike the studies in research communities, utility companies are the major entities that have access to tremendous amount of smart meter data. They are eager to utilize the data to know more about their customers. A typical example originates from marketing perspective, in which utility companies aim to detect a particular appliance. The reason is that many electric companies run various energy efficient marketing campaigns and try to encourage their customers replace inefficient appliances with more efficient alternatives. In general, marketing departments of the electric companies are required to run these campaigns through various communication channels such as bulk mail, direct phone calls, and so forth. However, the cost of reaching out to their large customer pool is significant. For instance, there are over 3 million household in Los Angeles, California. A rough cost per a bulk mail campaign is close to 1 USD (0.60 USD for a bulk mail and 0.40 USD for printing, sorting, and other labor cost), which costs more than 3 million USD for each marketing campaign. This type of campaign can run more efficiently if the electric companies could identify target customers who own and operate targeted appliances.

A fundamental functional block for the above example is an effective appliance detection algorithm. There are numerous methods that have been introduced by research communities [21, 28, 31]. However, the existing algorithms and approaches rely on special sensors at every device or a single sensor at power outlet [21]. For example, Weiss et al. [31] designed a system in which smart meter data sampled

at high frequency (1 hz), an appliance signature database and a single sensor monitoring the state of household current are integrated to disaggregate the total consumption into appliance level. Srinivasan et al. [28] applied several classifiers, e.g. Neural network and support vector machine to uniquely identify various types of devices using their distinct harmonic "signatures". All these works require data with a high sampling rate and prior domain knowledge of appliance signatures.

Unfortunately, it is nontrivial to adopt the existing approaches to solve the emerging appliance detection problem that the utility companies encounter. On one hand, they may not have fine grained data and appliance signatures. Given fine grained data, a simple clustering based approach can achieve very good performance [10]. However, given very coarse grained data such as daily consumption, it is very challenging to extract relevant features for building high quality and meaningful clusters hence utility companies have to design a more sophisticated approach. On the other, fully labeled data and/or controlled experimental setting may not be available in real-world applications. In other words, the available data to the utility companies is mostly partially labeled, which prevents them from adopting the state-of-the-art supervised algorithms, such as support vector machine (SVM) [30] and random forest [5].

To alleviate the requirement of high sampling frequency, Kolter et al. [14] used hourly consumption data and utilized discriminative sparse coding to model each device's power consumption over one week, then combine learned models to predict the power consumption at device level to unseen households. Therefore, the device categories, such as TV, refrigerators and electrical water heaters, can be inferred from different consumption level. But this approach assumes the individual device readings are available in the training phase, which is impractical in most cases. Besides, the granularity of data is still relatively high compared with daily data, which is available to vast population.

In this paper, we propose a novel approach to detect a particular appliance: "heat pump"[1] from coarse grained data with daily consumption and partial labeling information. More specifically, we utilize low granularity smart meter data, prior sales data and weather data to extract relevant features related with heat pump usage. Our approach does not have any assumption about the access to individual device consumption or the prior knowledge of any appliance signature, but correlates the daily consumption with temperature and extracts features specific to heat pump usage. Given the fact that only a small portion of heat pump users are available from prior sales record (a.k.a partially positive labeled data), such a problem is characterized by positive and unlabeled data learning, in which "positive" represents the small portion of users with heat pumps and "unlabel" means the rest large population that may or may not have heat pumps. We applied the state-of-the-art learning algorithm, biased SVM [18] that is designed specifically for positive and unlabeled data classification to our extracted features and achieved very competitive performance.

To summarize, our contributions in this paper are multifaced. At the conceptual level, we take the first step to detect a particular appliance from coarse grained data compared with other methods which depend on fine grained

data. At the system level, we integrate different data sets from different sources, including smart meter data, prior sales record and weather data, to perform feature extraction and heat pump detection. At the modeling level, we formulate the problem as a positive and unlabeled data classification problem and apply the state-of-the-art method to our extracted features, including temperature dependent features related with heating and wavelet features from heating period (typically from November to February). To our best knowledge, we present the first case of performing a particular appliance detection at a very coarse granularity smart meter data by integrating data sets from different sources. Experimental results show that our approach significantly improves prediction accuracy using real world data sets.

The rest of the paper is organized as follows: we first introduce related works in Section 2. We formalize the problem, and describe the details of our solution in Section 3. We present the experimental results in Section 4, and finally conclude the paper.

2. RELATED WORK

There are two research areas which are closely related to our work: appliance deteciton/recognition and positive and unlabeled learning.

2.1 Appliance Detection

The research topics on appliance detection problem can be classified into two categories based on the difference of hardware and software deployment. One is using single sensor to monitor the power outlet [21] or multiple sensor to measure the electrical information of individual appliance [27]. The major problem is that it requires special technical expertise for single sensor deployment or incurs huge expense for multiple sensor installation and maintenance.

The other utilizes data mining techniques to disaggregate overall consumption data into individual device level and associate different levels of consumption to existing appliance feature database [14, 24, 28]. The consumption data is typically sampled at high frequency, e.g. 1 second and 15 minutes. For example, Srinivasan et al. [28] extracted features from the input current waveform and applied several classifiers, such as Neural network and support vector machine to uniquely identify various types of devices. Kolter et al. [14] developed discriminative sparse coding to learn a model for the power consumption of each device. Therefore, the device categories, such as TV, refrigerators and electrical water heaters, were inferred from different power consumption level. However, the prior knowledge about appliance consumption is difficult to obtain due to the rapid changing world. In addition, hidden factors such as weather that affect energy consumption were ignored.

Our work is different from previous researches in the following sense. First, instead of detecting all appliances from a very fine granularity data, we detect a particular appliance: "heat pump", which is a major energy consuming appliance in residential household. Second, we do not assume any prior knowledge about the appliance signature, e.g. power consumption, on or off state current et al. Hence our method is totally data driven without incurring extra cost. Finally, we integrate different data sources into a unique system, in which sales record and weather data are utilized to guide more effect heat pump detection.

2.2 Positive and Unlabeled Learning

Positive and unlabeled learning (PUL) is extensively studied in data mining and machine learning communities from different perspectives, e.g. text categorization [18], Bioinformatics [9], Cheminformatics [34], collaborative filtering [22] et al.

Different from traditional supervised and semi-supervised learning where both positive and negative data are specified, the training data of PUL are composed of a set of positive data and a large amount of unlabeled data which can be positive or negative. Considering the heat pump detection problem, we only know a small portion of users having heat pumps from sales record data. For the rest large amount of users, they may or may not have heat pumps. Hence PUL is suitable tool since it aims to fully exploit the unlabeled data together with the limited positive data to learn more precise predictive models.

Two general approaches of PUL have been proposed. One is a two-step approach [19, 32], in which a certain reliable negative samples are iteratively identified from the unlabeled data first and then traditional classifiers (e.g. SVM, Naive Bayes) are applied to the reliable negative set and positive set. However, the performance of such a two-step approach highly depends on the quality of the identified negative samples. The other is a one-step approach [9, 18], in which all the unlabeled samples are treated as negative and the model is trained only once. For example, biased SVM [18] is obtained by introducing different misclassification cost of the positive and negative samples to ordinary SVM [30]. The underlying principle is that if the sample size is large enough, minimizing the number of unlabeled examples classified as positive while constraining the positive examples to be correctly classified will give a good classifier. Another example is logistic regression for positive and unlabeled learning (LRPU) [9], which estimates the conditional probability of the positive class given input samples directly. Though the marginal probability of the positive class and the conditional probability of the labeled positive samples have to be estimated as an intermediate step [35], it provides competitive performance as biased SVM as studied in [9].

In this paper, we adopt biased SVM (BSVM) [18] for heat pump detection after extracting features from daily consumption and weather data. Our experimental studies shows that BSVM outperforms LRPU on a real-world data set.

3. METHODOLOGY

In this section, we describe the proposed heat pump detection framework. As mentioned earlier, our framework can be divided into two phases: feature extraction and heat pump classification. Before introducing them in detail, we first outline the notation of this paper.

We use lowercase letters to represent scalar values, lowercase letters with bold font to represent vectors (e.g. β), uppercase letters to represent matrices, and uppercase calligraphic letters to represent sets. Unless stated otherwise, all vectors in this paper are column vectors.

3.1 Problem Statement

Given a smart meter data set $\mathcal{P}$ with daily consumption from known heat pump users[2] as well as another set of daily consumption data $\mathcal{U}$ from unknown users over the same time period, our aim is to build a predictive model that detects heat pump users.

3.2 Feature Extraction

The energy consumption data over a time period can be naturally modeled as a one-dimensional time series. Without loss of generality, assume $\mathbf{x} \in \{\mathcal{P} \cup \mathcal{U}\}$ is the consumption over discrete time stamps $t_1, t_2, \cdots, t_n$ for one particular user, then x_i is the consumption at time t_i.

A naive way to learn with time series data is to treat the value at each time point as a feature. However, a major problem of such a way is that the dimensionality of feature space is very high. The high dimensionality will introduce the curse of dimensionality and cause problems in distance metrics [4]. Fortunately, consecutive values of a time series are usually dependent, highly correlated and contain a lot of redundancy. Therefore we seek a good feature representation of time series that not only speeds up the learning algorithm, but has better performance.

First of all, we seek to choose some key empirical features that can capture key characteristics of the heat pump detection problem. The key underlying assumption is as follows. If there is an electric heating and cooling system in a building, daily average electric energy consumption will be higher if the outdoor temperature is low, or the indoor set temperature is high, or the building has high heat loss, and so forth. Our potential empirical feature candidates are: temperature dependent heating parameter, temperature dependent cooling parameter, the ratio between the average energy consumption during the cooling period [3] and the average energy consumption during non-cooling and non-heating period, and the average energy consumption during the heating period [4] and the average energy consumption during non-cooling and non-heating period. We will discuss how to calculate them empirically based on energy consumption in the next subsection.

In addition, we use generic features of time-series data to capture additional characteristics between heat pump and non-heat pump users. Popular feature extraction techniques for time series include the Discrete Wavelet Transform (DWT) and the Discrete Fourier Transform (DFT). Both transformation methods divide up time series data into different frequency components and then study each component with a resolution matched to its scale [17]. The main difference is that wavelets are localized in both time and frequency whereas the standard Fourier Transform is only localized in frequency. Another difference is that DWT is less computationally complex with $O(n)$ time as compared to $O(n log n)$ for the fast Fourier transform, where N is the length of time series. In this paper, we adopt the DWT not only because DWT is fast, but DWT has produced competitive or better results in a bunch of time series data mining tasks, such as clustering [12, 33], classification [8, 29] and similarity search [23, 16]. A more comprehensive survey can be found in [17].

3.2.1 Temperature Dependent Heating Features

Temperature dependent heating features include average consumption in the heating period, ratio between the average energy consumption during the heating period and the

---

[2]either from prior sales record or other reliable sources

[3]June to September

[4]November to February

average energy consumption during non-cooling and non-heating period and temperature dependent heating parameter. The first two are obvious hence we introduce the third term and explain why we focus on heating related features rather than cooling.

Space heating and cooling systems run to maintain indoor temperature at desired temperature levels. An electric heat pump system heats up and cools down a space with a bi-modal heat exchange mechanism. When outdoor temperature is lower than indoor temperature, heat travels through the building envelop, walls, windows, and ceilings to outside. This heat is lost by conduction. Also, heat travels through leaks, which is called infiltration. When outdoor temperature is higher than indoor temperature, heat travels into the building by conduction, infiltration, radiation, people activities, and appliances activities. Roughly speaking, indoor temperature is settled when (1) heat gain and transferred heat from indoor to outdoor are balanced in summer during summer and (2) heat loss and supplied heat are balanced during winter.

Once a building is built, building characteristics such as building envelop, walls, windows, ceilings, and openings do not change significantly over time. Only the outdoor temperature changes significantly day-by-day. Both conductive and infiltrative heat loss and gain are proportional to temperature difference between two thermal mass, which means that the space heating and cooling system consumes more energy when outdoor temperature is lower during winter and outdoor temperature is higher during summer.

Figure 1 illustrates the drybulb temperature [5] v.s. electric energy consumption of a building where an electric heat pump is installed. As we can see that there is a clear correlation between the outdoor temperature and electric energy consumption. In Figure 1, a and b are the heating and cooling slopes, which represent cooling and heating energy requirement per temperature drop (kWh/degree).

This correlation is used to quantify heating and cooling efficiency of a building and to predict heating and cooling energy consumption. One of the widely used model to identify this correlation is a piece-wise linear regression model [2, 25], which can be written as follows:

$$\begin{aligned} J &= aT + J_{bh} + \gamma & \text{when } T \leq T_h \\ &= J_b + \gamma & \text{when } T_h < T < T_c \quad (1) \\ &= bT + J_{bc} + \gamma & \text{when } T_c \leq T \end{aligned}$$

where J is a daily energy consumption in kWh, T is drybulb outdoor temperature, J_{bh} is nominal energy consumption during heating required days, J_{bc} is nominal energy consumption during cooling required days, J_b is nominal energy consumption during heating required days, γ is a random variable that has a certain probability distribution with bounded second moment, a and b are heating and cooling consumption slopes.

In buildings where there is no electric heating nor cooling system, a and b are close to zero whereas in buildings with electric heat pump systems, a is a noticeable negative slope and b is a positive slope. In this paper, we also call the

---

[5]drybulb temperature is the temperature measured freely in the air without radiation and moisture, which is considered to be the most important variable for building energy loss and gain
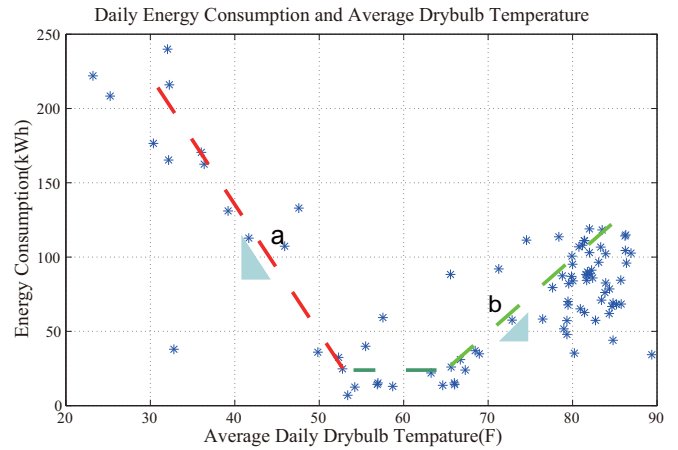
Figure 1:   Energy Consumption in a House with an Electric Heat Pump System
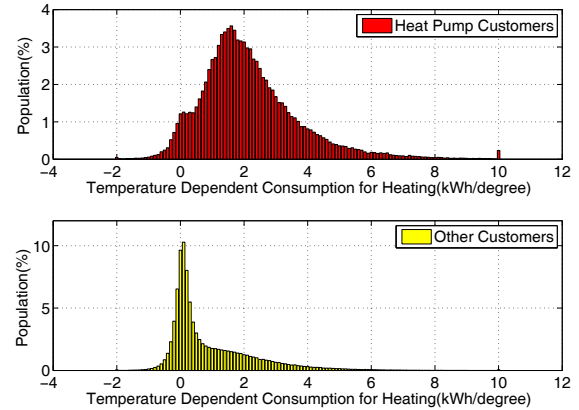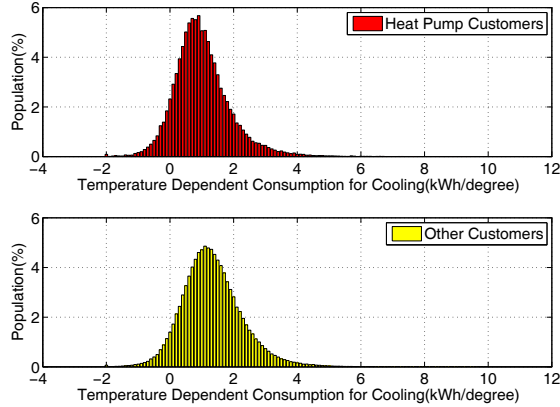

Figure 2:   Histogram of Heating Slopes: Customers with electric heat pump and others

heating slope a as *temperature dependent heating parameter* and b as *temperature dependent cooling parameter*.

A natural question is whether we should use both a and b for prediction. Intuitively, high consumption in summer or winter may not be from heat pump only. Other appliances, such as windows AC unit and electric heater, can have correlation between temperature and energy consumption, though the slope may not be steep. To validate the effectiveness of them, we plot the distribution of a and b in Figures 2 and 3 with two sets of customers who own electric heat pumps and other systems. Figure 2 shows a clear distinction between the electric heat pump customers and other customers. One possible reason is that majority of space heating systems use gas or heating oil, thus the customers who do not own an electric heat pump system tend to consume more gas or oil other than electricity.

It is worthwhile to inspect whether the temperature dependent cooling parameter b has similar characteristics to a. Figure 3 shows the histograms of the temperature dependent cooling parameter. As we can see, two groups have similar distributions. One possible reason is that majority of

Figure 3: Histogram of Cooling Slopes: Customers with electric heat pump and others

Table 1: Haar wavelet transform on a four element time series with different levels.

Level	Approximations	Detail coefficients
0	8,6,2,10	
1	7,6	-1,4
2	6.5	-0.5

Haar transformation, one can fully recover the original signal [17].

Figure 4: Tree structure of two-level Haar transformation. x is the input time series, AX_i is the ith level approximation and DX_i is the detail coefficients.

non-heat pump users use electricity based cooling systems, e.g. windows AC unit. Though without heat pump usage, the total energy consumption still have high correlation with drybuld temperature, which makes the temperature dependent cooling parameter b useless to distinguish between heat pump users and non-heat pump users. Based on the comparison between Figure 2 and 3, we conclude that the temperature dependent heating parameter a is more useful. Hence, we only extract heating related features.

In addition to the temperature dependent heating parameter a, we also collect the average consumption in the heating period, the ratio between the average energy consumption during the heating period and during non-cooling and non-heating period. These two parameters are added to deal with electric heat pump users who have high variability in energy consumption during the heating period.

3.2.2 Wavelet Features

Though these temperature dependent heating features have a certain discriminative power to differentiate between heat pump and non-heat pump users, we argue that they can not capture all the characteristics of heat pump usage, especially in the time and frequency domain. Therefore, we still need to extract a set of generic features from energy consumption time series.

As discussed previously, we utilize discrete wavelet transform (DWT) to extract generic features. In this paper, we use Haar transform, which is a simple but powerful mother wavelet functions. Haar transform can be viewed as a series of averaging and differencing operations on the time series. The averaging part captures the trend of time series called approximation, and the differencing part depicts the surprise named detail coefficients. To illustrate the procedure, we compute the averages and differences between every two adjacent values of $\mathbf{x} = (8, 6, 2, 10)^T$ as shown in Table 1. The level 0 gives the original time series. In level 1, $(7, 6)$ are obtained by taking the average of $(8, 6)$ and $(2, 10)$ respectively. $(-1, 4)$ are the differences of $(8, 6)$ and $(2, 10)$ divided by 2 respectively. This process is repeated until level 2 (only one approximation coefficient left) is reached. The final Haar transform $H(x) = (6.5, -0.5, -1, 4)$ is obtained. From the
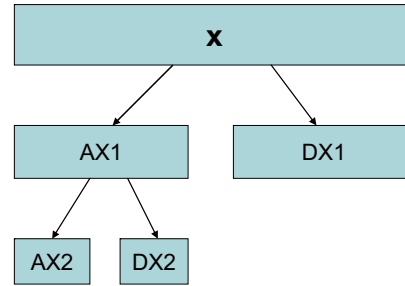
The multi-level decomposition can be conveniently modeled as a tree shown in Figure 4, in which the root mode represents the original time series, and left/right child represents approximations/detail coefficients respectively. Note that we just show two-level decomposition, there may be even more deeper decomposition for high dimension time series. Suppose kth level DWT produces the best performance on validation set, then the transformed feature vector from $\mathbf{x}$ is given by $(AX_k, DX_k, DX_{k-1}, \cdots, DX_1)$. k can be tuned by cross-validation and we fix $k = 2$ throughout the paper. Our empirical study shows that $k = 2$ gives the best performance.

Instead of applying DWT on the whole time period including four seasons, we specifically consider two specific seasons: summer and winter. The reason is that the two seasons require heating or cooling, which triggers a lot of consumption as a result of heat pump usage. As discussed beforehand, there may be a lot of consumption in cooling period, but it may be as a result of central AC or window AC unit. Therefore, the consumption in cooling period is not quite useful for determining heat pump usage. Similarly, for the time period without cooling or heating, there is no or seldom heat pump usage and the consumption is useless as well. Considering the above factors, we only extract wavelet features from heating period, e.g. November to February.

3.3 Heat Pump Classification

Once features are extracted from heat pump users and unknown users, the problem is converted to a classification problem with positive and unlabeled data. To tackle the problem, we utilize the state-of-the-art algorithm biased SVM [18].

Followed the convention in [18], suppose training sets are $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2),$
$\cdots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i$ is the ith input vector and y_i is

its class label, $y_i \in \{1, -1\}$. Assume that the first $l - 1$ examples are positive examples (labeled 1), while the rest are unlabeled examples, which we label as -1. It was shown in [19] that if the sample size is large enough, minimizing the number of unlabeled examples classified as positive while constraining the positive examples to be correctly classified will give a good classifier [18]. In particular, consider the following objective:

$$\min_{\mathbf{w}, b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_p \sum_{i=1}^{l-1} \xi_i + C_u \sum_{i=l}^{n} \xi_i$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) >= 1 - \xi_i, \ i = 1, 2, \cdots, n \quad (2)$$
$$\xi_i >= 0, \ i = 1, 2, \cdots, n$$

where C_p and C_u are regularization parameters to control the fitness for positive and unlabeled samples. We can vary C_p and C_u to enforce which part to be correctly classified. Intuitively, we give a big value for C_p and a small value for C_u because the unlabeled set, which is assumed to be negative, also contains positive data. The objective (2) is convex and a unique global solution exists. In this paper, we use the standard convex optimization package cvx [11] to solve the objective.

To choose C_p and C_u, a typical approach is to use a separate validation set or cross validation to verify the performance of the resulting classifier with the selected values for C_p and C_u. A widely used metric is the F_1 score defined as $F_1 = 2pr/(p + r)$ where p is the precision and r is the recall.

Unfortunately, it is impossible to evaluate precision without knowing negative examples. Hence we follow the psudo-F_1 metric:

$$psudoF_1 = r^2/Pr(f(\mathbf{x}) = 1) \quad (3)$$

which is proposed in [15], where $Pr(f(\mathbf{x}) = 1)$ is the probability that a user is classified as heat pump user. r can be estimated using the positive examples in the validation set and $Pr(f(\mathbf{x}) = 1])$ can be estimated from the whole validation set. As explained in [18], this criteria works because it behaves similarly to the F_1 score in the sense that it is large when both p and r are large and is small if either p or r is small.

4. EXPERIMENT

We have conducted a rigorous evaluation of our method in terms of detection accuracy using a set of real-world data, including smart meter consumption data, prior sales record data and local weather data collected from an anonymous utility company [6]. We implemented a prototype of our method in Matlab and compared our method with other two state-of-art models including ordinary Support Vector Machine (OSVM) [30], logistic regression for positive and unlabeled data classification (LRPU) [9].

4.1 Data sets

To evaluate our method, we utilized a smart meter data set (daily consumption from 01/01/2011 to 09/23/2012) and a heat pump sales data set from an anonymous utility company. In the data set, there are around 300k users residing in 6 regions. Based on the sales record, we single out 4565 electrical heat pump users who purchased heat pump in 2009

or 2010 and 1821 non-heat pump users before March 2012 [7]. We also obtained the weather data for each region in the same time period as smart meter data. For evaluation purpose, we treat all 1821 non-heat pump users as unlabeled samples and randomly sample 10% (456) heat pump users named Q as additional unlabeled samples. Let P be positive samples and U be the unlabel samples and N be the negative samples, it is clear that P has the cardinality of 4100 and $U = N \cup Q$ has the cardinality of 2277.

As described in Section 3, our method extracts features specific to heating. The features include temperature dependent heating parameter, average daily consumption during heating period, the ratio between average consumption during heating period and other time periods other than heating or cooling period and 2-level wavelet transform coefficients from the time series of consumption in heating period. There are two heating periods given the time span, including 01/01/2011-02/28/2011 and 11/01/2011-02/28/2012. We ignore the consumption after March 2012 since the negative set is valid only before that time and the total number of features is 139.

Besides the small data set with ground truth that can be used for model evaluation, we also have a large smart meter data set with 292,496 users without labels. Since there is no ground truth, we only provide prediction result and match any existing knowledge about heat pump sales or market share.

4.2 Experimental Protocol

We use standard 10 fold cross validation to generate training and testing data sets. For ordinary SVM, we use the libsvm [6] package with linear kernel. The C parameter is tuned from $1, 10, 20, 30 \cdots, 100$. For biased SVM, there are two parameters C_u and C_p. The first parameter, C_u, is to control the fitness of unlabeled set. The second parameter, C_p, is to control the fitness on positive labeled set. As suggested by [18], we choose C_p from a set of larger numbers than C_u. More specifically, we tune C_p from $20, 25, 30, \cdots, 50$ and C_u from $1, 3, 5, \cdots, 17$. Below we summarize the model construction and model evaluation.

Model Construction and Selection. For each data set, we partition the data set into 10-folds to perform 10-fold cross-validation (CV) with 9 folds for training and 1 fold for testing. We use an internal 5-fold CV on the training data set to select the optimal parameters for ordinary SVM and biased SVM based on the pseudo-F1 score defined in 3. We then generate a single model from the entire training set with the selected parameters and apply the model to the testing data set for prediction. For logistic regression proposed in [9], there is no additional parameter to tune hence the model is obtained from entire training data without interval cross validation. Furthermore, all methods treat positive labeled samples as positive and unlabeled data as negative.

Model Evaluation. For model evaluation, we collect the precision: $(TP/(TP + FN))$, recall: $(TN/(TP + FP))$, F_1 score: $(2*\text{precision}*\text{recall}/(\text{precision}+\text{recall}))$ and accuracy: $((TP + TN)/S)$ of the trained model. Note that since the "real positives" are positive samples and a few unlabeled samples in test set, the confusion matrix is a little different from binary classification. For each trial i, $i = 1, 2, \cdots, 30$, a classifier is applied to the test set and yield the following

---

[6]Due to the confidentiality agreement, the company's information is anonymized.

[7]They purchased heat pumps not for replacement after March 2012
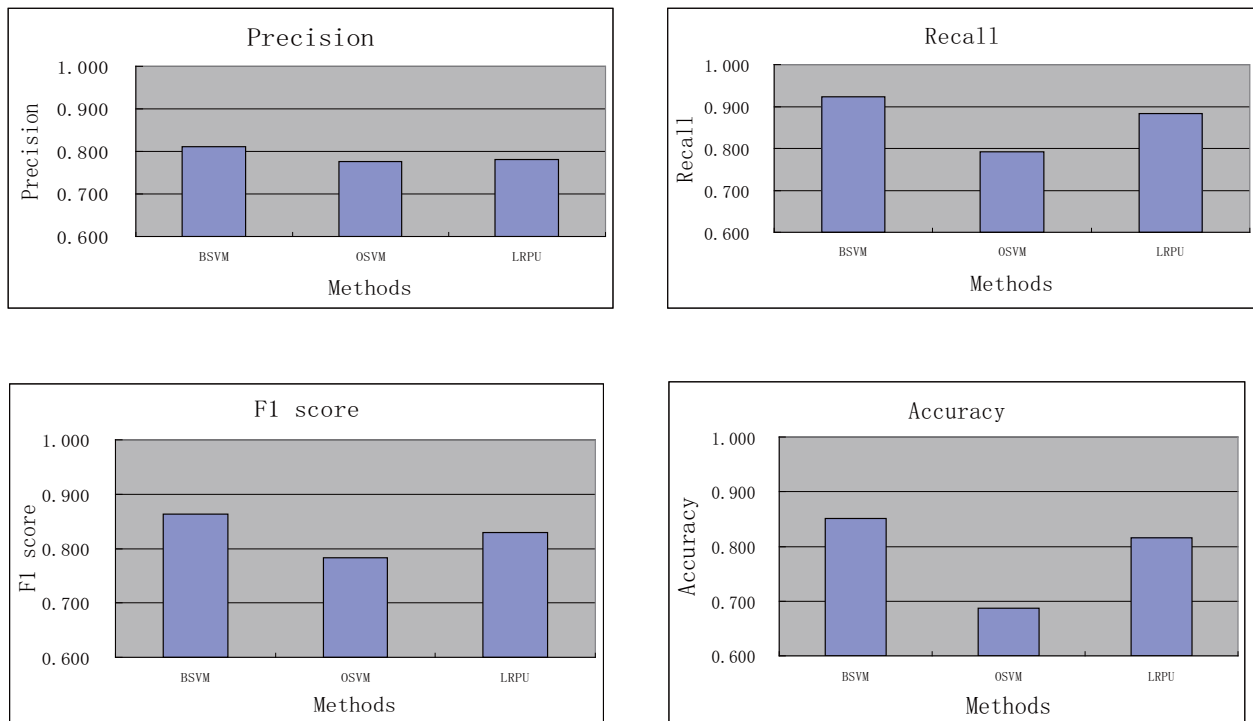
Figure 5: Performance comparison over three methods. Top left: precision; top right: recall; lower left: $F1$ score; lower right: accuracy.

confusion matrix: where TP stands for true positive, FP

Table 2: The confusion matrix for one cross validation trial. P_i, Q_i and N_i are subsets from P, Q and N respectively.

		Predicted	
		positive	negative
Actual	$P_i \cup Q_i$	TP	FN
	N_i	FP	TN

stands for false positive, TN stands for true negative, FN stands for false negative, and S stands for the total number of samples.

All the values reported are collected from the testing data set only and are averaged across 3 replicates of the 10-fold cross validation in a total of 30 experiments.

4.3 Classification Performance

In this subsection, we show the performance of biased SVM (BSVM) compared with ordinary SVM (OSVM) and logistic regression for positive and unlabeled data mining (LRPU) [9]. The precision, recall, $F1$ score and accuracy is shown in Figure 5. Since the standard deviation is around 1%-2% for each method, we omit them for simplicity.

From the lower right subfigure in Fig 5, we observe that the accuracy of ordinary SVM with linear kernel performs the worst. It is understandable since ordinary SVM imposes the same penalty on both positive and "negative" data set but the "negative" data set is not purely negative. Comparing BSVM with LRPU, BSVM performs slightly better

than LRPU. The possible reason is that biased SVM focuses predicting positive samples correctly while controlling the number of samples classified as negative. When the data set has many more positive samples, BSVM outperforms LRPU, though the latter approach has one step of probability adjustment in training phase. To the contrary, LRPU performs slightly better when the data set has more underlying negative samples. More detailed comparison between BSVM and LRPU can be found in [9].

To better understand the accuracy differences, we plot the average precision, average recall and average F_1 score of all methods in Fig 5. Similar to accuracy, OSVM performs worst in terms of precision, recall and F_1 score. The performance of LRPU is a little inferior to BSVM for precision and F_1 score, though the performance gap is not that significant. BSVM has a very high recall (a.k.a. the algorithm predicts most positive samples correctly) and a reasonable precision, which is supportive to the analysis beforehand. Such a performance of BSVM is desirable in a few applications such as heat pump marketing campaigns, because utility companies may not want to target too many customers who have heat pumps already.

4.4 Feature Extraction Evaluation

We have demonstrated the classification performance of three classifiers on our extracted features. A natural question is how useful of the heating specific features. To validate the usefulness of extracted features, we apply biased SVM (BSVM) classifier to our features (HFeature) and general wavelet features (GFeature) extracted from all time period up to March 12.

In Table 3, it is clear that BSVM with our extracted features (HFeature) is superior to general wavelet coefficient features. For simplicity, we do not report standard deviation since it is around 1%-2% for each feature set. The general features (GFeature) that are extracted from all time periods not only contain the heating period, but encompass cooling and no cooling/heating period, in which the consumption is either from AC system for cooling or regular appliance without heat pump usage. Therefore, it is difficult to obtain an accurate classifier from such a feature set.

Table 3: Performance comparison between our extracted features (HFeature) and general features (GFeature).

	F_1	Precision	Recall	Accuracy
HFeature	0.864	0.811	0.923	0.852
GFeature	0.785	0.758	0.813	0.716

4.5 Prediction on the large data set

As discussed before, the large data set with 292,496 users has no labeling information. The best result that we can deliver with this data set is to run our algorithm on it and match the result with existing knowledge of heat pump sales or market share. We trained a BSVM model from all the 4565 users applied it to the large data set.

Although rigorous assessment of the performance of proposed algorithm cannot be conducted on this data set, two evidences show the proposed algorithm work reasonably well. First, the proposed algorithm identified 129,238 (44.2%) as heat pump users from the 292,496 users. The result is consistent with the market share of heat pump in that region [1]. In addition, the classification result has been used by the partner electric company which had an existing algorithm to identify electric heat pump users. During the pre-campaign phase, the partner company randomly singled out 20 customers who are predicted as non-heat pump users by our algorithm for evaluation purpose. Among the 20 customers, 12 customers were identified as heat pump users by their algorithm, which is simply based on the overall energy consumption and the correlation with temperature. To validate the effectiveness of two models, We went through a validation process and found that the result of our method was indeed more accurate than the existing algorithm of our partner company. All the 12 customers are "real" non-heat pump users.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a heat pump detection method for a targeted energy efficient marketing campaign. This simple yet practical problem has two unique challenges: (1) the input data is only partially labeled with only a subset of positive samples; and (2) only the extremely coarser grained energy consumption data is available for the detection algorithm due to the current smart meter reading infrastructure.

To tackle the two challenges, we formulate the detection problem as a positive and unlabeled data learning problem and adopt biased SVM to solve it. Besides, we extract both empirical features and generic features relevant to heat pump usage in heating period for classification. We

have evaluated the performance against 4565 users' data and showed our approach achieves better accuracy than other competitive methods.

Our method has been implemented and deployed in a real-life setting where the partner electric company runs this targeted campaign for 292,496 customers without prior sales record. Based on the initial feedback, our detection algorithm can successfully detect substantial number of non-heat pump users who were identified as heat pump users with the prior algorithm the marketing department had used.

The future work will focus on two directions: (1) create a good user interface to make the analytics easier for utility companies; (2) improve the scalability of the method. For example, for this simple task, the analytics needed to process approximately 970 million data points for approximately 300k customers and run a couple of hours in a single machine. This number can be easily over multi-million customers served by a single grid operator. Therefore, the scalability and latency consideration of the algorithm need to be improved.

6. REFERENCES

[1] Ground source heat pumps: Overview of market status, barriers to adoption, and options for overcoming barriers. Technical report, Navigant Consulting, Inc., 2009.

[2] O. Balaca, H. Bulut, and T. Yilmaz. Analysis of variable-base heating and cooling degree-days for turkey. *Applied Energy*, 69(4):269 – 283, 2001.

[3] C. Bandim, J. Alves, J.E.R., J. Pinto, A.V., F. Souza, M. Loureiro, C. Magalhaes, and F. Galvez-Durand. Identification of energy theft and tampered meters using a central observer meter: a mathematical approach. In *Transmission and Distribution Conference and Exposition, 2003 IEEE PES*, volume 1, pages 163 – 168, sept. 2003.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *In Int. Conf. on Database Theory*, pages 217–235, 1999.

[5] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[6] C. Chang and C. Lin. Libsvm: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[7] S. chiang Lee, G. yuan Lin, W. rong Jih, and J. Y. jen Hsu. Appliance recognition and unattended appliance detection for energy conservation. In *Plan, Activity, and Intent Recognition*, 2010.

[8] G. V. Dijck, M. Wevers, and M. M. V. Hulle. Corrosion Time Series Classification using the Continuous Wavelet Transform and MML Density Estimation. In *IJCI-IJIT-IJSP Conferences*, pages 39–43, 2004.

[9] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 213–220, New York, NY, USA, 2008. ACM.

[10] H. Goncalves, A. Ocneanu, and M. Berges. Unsupervised disaggregation of appliances using aggregated consumption data. In *1st KDD Workshop*

on *Data Mining Applications in Sustainability (SustKDD)*, Aug. 2011.

[11] M. Grant and S. Boyd. *CVX: Matlab software for disciplined convex programming*, December 2008. Web page and software available at `http://stanford.edu/~boyd/cvx`.

[12] H. Guo, Y. Liu, H. Liang, and X. Gao. An application on time series clustering based on wavelet decomposition and denoising. In *Natural Computation, 2008. ICNC '08. Fourth International Conference on*, volume 5, pages 419 –422, oct. 2008.

[13] T. Kato, H. S. Cho, D. Lee, T. Toyomura, and T. Yamazaki. Appliance recognition from electric current signals for information-energy integrated network in home environments. In *Proceedings of the 7th International Conference on Smart Homes and Health Telematics: Ambient Assistive Health and Wellness Management in the Heart of the City*, ICOST '09, pages 150–157, 2009.

[14] J. Z. Kolter, S. Batra, and A. Y. Ng. Energy disaggregation via discriminative sparse coding. In *NIPS*, pages 1153–1161, 2010.

[15] W. S. Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*, pages 448–455, 2003.

[16] Y. leh Wu, D. Agrawal, and A. E. Abbadi. A comparison of dft and dwt based similarity search in time-series databases. In *In Proceedings of the 9 th International Conference on Information and Knowledge Management*, pages 488–495, 2000.

[17] T. Li, Q. Li, S. Zhu, and M. Ogihara. A survey on wavelet applications in data mining. *SIGKDD Explorations*, 4(2):49–68, 2002.

[18] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *ICDM*, pages 179–188, 2003.

[19] B. Liu, W. S. Lee, P. S. Yu, and X. Li. Partially supervised classification of text documents. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pages 387–394, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[20] D. Mashima and A. A. Cárdenas. Evaluating electricity theft detectors in smart grid networks. In *RAID*, pages 210–229, 2012.

[21] F. Mattern, T. Staake, and M. Weiss. Ict for green: how computers can help us to conserve energy. In *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking*, e-Energy '10, pages 1–10, New York, NY, USA, 2010. ACM.

[22] R. Pan, Y. Zhou, B. Cao, N. Liu, R. Lukose, M. Scholz, and Q. Yang. One-class collaborative filtering. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 502–511. IEEE, 2008.

[23] I. Popivanov and R. J. Miller. Similarity search over time-series data using wavelets. In *ICDE*, pages 212–221, 2002.

[24] J. Powers, B. Margossian, and B. Smith. Using a rule-based algorithm to disaggregate end-use load profiles from premise-level data. *Computer Applications in Power, IEEE*, 4(2):42 –47, april 1991.

[25] M. T. R Lindberg, A Binamu. Five-year data of measured weather, energy consumption, and time-dependent temperature variations within different exterior wall structures. *Energy and Buildings*, 36:495–501, 2004.

[26] T. Saitoh, Y. Aota, T. Osaki, R. Konishi, and K. Sugahara. Current sensor based non-intrusive appliance recognition for intelligent outlet. In *The 23rd International Technical Conference on Circuits/System, Computers and Communications*, 2008.

[27] H. Serra, J. Correia, A. Gano, A. de Campos, and I. Teixeira. Domestic power consumption measurement and automatic home appliance detection. In *Intelligent Signal Processing, 2005 IEEE International Workshop on*, pages 128 – 132, 2005.

[28] D. Srinivasan, W. Ng, and A. Liew. Neural-network-based signature recognition for harmonic source identification. *Power Delivery, IEEE Transactions on*, 21(1):398 – 405, jan. 2006.

[29] A. Subasi. Eeg signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications*, 32(4):1084 – 1093, 2007.

[30] V. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.

[31] M. Weiss, A. Helfenstein, F. Mattern, and T. Staake. Leveraging smart meter data to recognize home appliances. In *PerCom'12*, pages 190–197, 2012.

[32] H. Yu, J. Han, and K. C.-C. Chang. Pebl: positive example based learning for web page classification using svm. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 239–248, New York, NY, USA, 2002. ACM.

[33] H. Zhang, T. B. Ho, Y. Zhang, and M. song Lin. Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *INFORMATICA*, 2006.

[34] Y. Zhao, X. Kong, and P. Yu. Positive and unlabeled learning for graph classification. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, Dec. 2011.

[35] J. T. Zhou, S. J. Pan, Q. Mao, and I. W. Tsang. Multi-view positive and unlabeled learning. *Journal of Machine Learning Research - Proceedings Track*, 25:555–570, 2012.