Preprints of the 10th IFAC International Symposium on Dynamics and Control of Process Systems
The International Federation of Automatic Control
December 18-20, 2013. Mumbai, India

Optimal Feature Selection for SVM Based Fault Diagnosis in Power Transformers

Mahak Mittal * **Mani Bhushan** * **Shubhangi Patil** ** **Sushil Chaudhari** **

* *Department of Chemical Engineering, Indian Institute of Technology Bombay, India 400076 (Email: mahakmittal3@gmail.com, mbhushan@iitb.ac.in)*
** *High Voltage Product Technology Center, Global R &D, Crompton Greaves, Mumbai, India 400076 (Email: shubhangi.patil@cgglobal.com, sushil.chaudhari@cgglobal.com)*

Abstract: Power transformer is one of the most vital equipment in an electrical system and its failure results in huge economic losses. Amongst the various data driven techniques available in literature for diagnosing faults in a power transformer, Support Vector Machine (SVM) is one of the most promising. In this context, SVMs have typically been implemented using all the gaseous species available from dissolved gas analysis (DGA). In this work, we propose to enhance the diagnostic performance of SVMs by using them with an optimally identified subset of gaseous species available from DGA. We propose to use mutual information to identify these optimal species (features). The approach is applied on industrial datasets corresponding to various commonly encountered faults in power transformers. The results show that better diagnostic performance is obtained when CO_2 concentration measurement is not used.

Keywords: Feature Selection, Mutual Information, Support Vector Machine (SVM), Power Transformer, Dissolved Gas Analysis (DGA)

1. INTRODUCTION

Power transformers are one of the most expensive piece of equipment used for power generation and transmission. Failure of power transformers can cause huge economic losses as the cost of replacement, transportation, installation and repairs are very high. Some of the commonly encountered faults associated with the power transformer include partial discharge, arcing, overheating and cellulose degradation (IEEEGuide, 2009). Dissolved Gas Analysis (DGA) is popularly used to diagnose these faults in oil-immersed transformers. DGA refers to the analysis of the gases dissolved in the oil bath of a transformer. These gases are formed due to degradation of oil and other insulating materials.

Various standards such as Key Gas Method, IEC ratios, Rogers Ratio (IEEEGuide, 2009; Duval, 1989; Rogers, 1978; Fist, 2000) etc. are used to interpret the DGA data for fault diagnosis. These standards are easy to use as they can be applied across a variety of power transformers. Additionally several data driven techniques such as PCA (Erdal et al., 2009), neural networks (Zhengwei et al., 2009), fuzzy logic (Naresh et al., 2008) and support vector machines (Lv et al., 2005; Bacha et al., 2012; Han et al., 2011) have also been applied in literature to the DGA data for diagnosing faults in power transformers. These data driven methods are specific to the available datasets (transformers) and hence are expected to yield better diagnostic performance than the available standards. Amongst the data driven methods, SVM is one of the most promising. Fault diagnosis systems for power transformers based on SVM have been shown to be more efficient when compared to other techniques such as ANN, fuzzy logic etc (Lv et al., 2005). SVM has several desired features, such as: it requires comparatively lower computational efforts while training, the training problem can be solved optimally, nonlinear decision boundaries can be easily obtained and it works well even when only a few samples are available.

It is well known that the diagnostic performance of data driven techniques can be enhanced by using an appropriately selected set of features. Of the several variables measured in a process, some may be non-informative for the diagnosis problem at hand (Verron et al., 2008). Use of these variables while performing fault diagnosis can lower performance due to noise associated with such variables. Feature selection techniques can be used to identify the most optimal set of features which can enhance the performance of a diagnostic technique. However, to the best of our knowledge, in the area of fault diagnosis of power transformers, the possibility of using feature selection techniques to enhance the performance of the SVM based diagnostic technique has not been investigated. The aim of our current work is to use an appropriate feature selection technique to improve the performance of the SVM based fault diagnosis in power transformers.

Several feature selection techniques are available in general pattern classification literature. In particular, feature selection based on mutual information proposed by Verron et al. (2008) is an effective way for selecting the optimal set of variables while developing a classifier. Mutual in-

formation quantifies the mutual dependence between the process variables and the class. For a specified number of variables, the set of variables giving the maximum mutual information can be identified as the most promising set of variables for efficient fault diagnosis.

In this work, we propose use of mutual information to select optimal set of features that lead to enhanced performance of SVM based diagnostic technique for power transformers, as compared to using all features. We apply our approach to industrial DGA datasets. In particular, the available data corresponds to different operating conditions (fault and normal) of power transformers. Thus the resulting fault diagnosis problem is a multi-category classification problem. We use multilayer SVM classifier to classify this multi-category data. Each layer of this multilayer SVM classifies the data to one of the two appropriately selected classes. For different number of specified features, the most promising set of features is obtained via mutual information. The classification accuracies of the resulting SVM classifiers for these various number of specified features are then compared to obtain the optimal set of features. The proposed approach of using mutual information based feature selection in combination with support vector machines as a classifier, can be applied to variety of other fault diagnosis problems as well.

The rest of the paper is organized as follows: in Section 2, we describe the relevant techniques used in the paper. In Section 3, we discuss the industrial datasets available to us, the architecture of the multilayer SVM classifier to be used to classify the available datasets and the proposed feature selection technique. The feature selection results are also reported and discussed in this section. The paper is concluded in section 4.

2. DESCRIPTION OF RELEVANT TECHNIQUES

In this section, we first discuss the Support Vector Machines in general followed by the description of the feature selection technique based on mutual information.

2.1 Support Vector Machines

Support vector machine (SVM) is a supervised learning method used in pattern classification and regression (Vapnik, 1995). In our paper, we summarize the classification application of SVM. An SVM can be used to classify data belonging to one of the two classes as illustrated in Figure 1. In this figure, the two symbols (triangle and diamond) represent the data belonging to two classes. Given a set of training samples as in Figure 1, a number of classifiers (boundary separating the two classes) can be obtained which separate the two classes. SVM involves finding an optimal hyperplane which maximizes the margin between the two classes. Support vectors are the data samples which are closest to the optimal hyperplane in both the classes and are thus most difficult to classify. The equation of the hyperplane is described with the help of these support vectors. Example of a linear classifier obtained using SVM is shown in Figure 1. In this figure, the solid line represents the linear hyperplane obtained by using SVM. The filled data points are the support vectors.

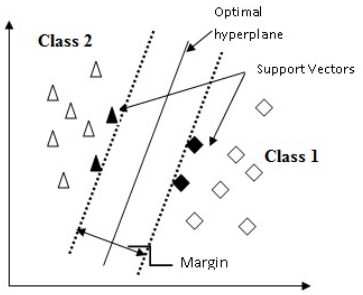

Fig. 1. Separation of two classes by linear SVM

In several situations, the data belonging to different classes overlaps and hence a linear hyperplane as in Figure 1 will not be able to completely separate the two classes. For such cases, nonlinear SVM can be used. In our work, we also propose to use nonlinear SVM since the fault classes in a power transformer are expected to be overlapping. The algorithm for obtaining such nonlinear SVM is thus described as follows:

Given the training dataset $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$ is i^{th} training sample and $y_i \in \{-1, 1\}$ denotes the corresponding class label. In nonlinear SVM, the data is mapped from d-dimensional space to N-dimensional feature space by employing a nonlinear function as (Lv et al., 2005):

$$\psi(x) = [\varphi_1(x) \;\; \varphi_2(x) \;\; \ldots \;\; \varphi_N(x)]^T \tag{1}$$

A linear SVM is then obtained in this high dimensional space. Based on this SVM, an observation vector x can be classified to one of the two classes depending on the value of class label $y(x)$ obtained as: $y(x) = sgn(w.\psi(x) + b)$ where $sgn(.)$ denotes the sign $(\pm 1, 0)$ function, with 0 indicating a tie. w is the weight vector and b is the bias corresponding to the linear SVM in the N dimensional feature space. These parameters are to be obtained so that the margin is maximized in the N dimensional features space. This can be achieved by solving the following optimization problem (Lv et al., 2005):

$$\left.\begin{aligned} &\min_{(w,b,\xi)} \; J = \tfrac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{such that,} \\ &y_i[w.\psi(x_i) + b] \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, n \\ &\xi_i \geq 0, \qquad i = 1, 2, \ldots, n \end{aligned}\right\} \tag{2}$$

In above formulation, ξ_i are the slack variables added to allow misclassification. The constant C, referred to as the regularization parameter, is user specified and controls the relative influence of the two competing terms (the margin and the misclassification). Instead of directly solving the above formulation, the dual of the above optimization problem is used to obtain the support vectors. The dual formulation is presented in equation 3 (Lv et al., 2005). The dual optimization formulation is quadratic programming in nature. If $a_i > 0$, then the corresponding x_i represents the support vector. The advantage of formulation 3 over formulation 2 is that in formulation 3 only the inner product of samples in the higher dimensional space needs to be known.

$$\left.\begin{aligned} \max_a W \quad &= -\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j (\psi(x_i).\psi(x_j)) + \sum_{i=1}^{n} a_i \\ \text{such that,} \quad & \\ & \sum_{i=1}^{n} a_i y_i = 0 \\ & 0 \le a_i \le C, \qquad \forall i = 1, 2, \ldots, n \end{aligned}\right\} (3)$$

Such inner products can be specified in terms of a Kernel function (K). Thus the optimum hyperplane equation can be obtained by using the following equation

$$\sum_{i \in SV} a_i y_i K(x_i, x) + b = 0 \tag{4}$$

where SV represents the set of indices of the support vectors. The bias b in Eq (4) is calculated using the following equation:

$$b = y_m - \sum_{i \in SV} a_i y_i K(x_i, x_m), \text{ where } m = \arg[\max_i a_i] \tag{5}$$

The selection of Kernel function $K(x_i, x_j) = \psi(x_i).\psi(x_j)$ which is a symmetric positive definite function in original space $\mathbb{R}^d$ based on Mercer conditions is an important design choice in SVM (Lv et al., 2005). Some of the popularly used Kernel functions are (Lv et al., 2005): Linear, Polynomial, Gaussian radial basis and Sigmoid. The Gaussian radial basis kernel given as

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|_2^2}{2\sigma^2}\right) \tag{6}$$

will be used in our work.

SVM for Multicategory Classification: The above was a summary of obtaining SVM for a two class classification problem. A classification problem involving several classes can be solved by using multiple SVMs where each SVM separates two appropriately selected classes. This architecture will be discussed in section 3 where we present the proposed work.

2.2 Feature Selection using Mutual Information

In this paper we use mutual information based feature selection technique as presented in Verron et al. (2008). It consists of two steps: i) sorting the variables according to the mutual information a variable shares with the class, and ii) selection of the best set of variables based on the classification accuracy. These two steps are summarized next:

Mutual Information: The mutual information $I(X, Y)$ of two random vectors X and Y can be viewed as a quantity measuring the mutual dependence between these two vectors and can be computed as (Verron et al., 2008):

$$I(X, Y) = \sum_{x,y} P(x, y) log \frac{P(x, y)}{P(x)P(y)} \tag{7}$$

In above equation, $P(x, y)$ is the joint probability mass function of X and Y, and $P(x)$ and $P(y)$ are the marginal probability mass functions. In supervised classification the classes (C) is viewed as a multinomial random variable with k possible values (k is the number of classes) and the probability mass function is given as P(C = c) = P(c). Here P(c) can be thought of as the prior probability of the data belonging to the c^{th} class. For computing the mutual information between the process variable vector X and the class the following are assumed: (i) X is a random variable with a multivariate normal density function i.e. $X \sim N(\mu, \Sigma)$ and (ii) X conditioned on C = c follows a multivariate normal density function with parameters μ_c and Σ_c. The mutual information between X and C then turns out to be (Verron et al., 2008):

$$I(X, C) = \frac{1}{2}\left[log(\mid \Sigma \mid) - \sum_{c=1}^{k} P(c) log(\mid \Sigma_c \mid)\right] \tag{8}$$

Mutual information can now be computed for all possible groups of available variables. The most important group of variables for the classification task will be the one that has a large mutual information. However, since mutual information will increase as more and more variables are considered, mutual information is used to select an appropriate set of variables when the number of variables p to be selected is specified.

Selection of the best group of features: In approach used by Verron et al. (2008), the classification accuracy using an appropriate classifier is computed for all the groups obtained for various values of p. The group corresponding to the maximum classification accuracy is then selected as the best set of features.

3. PROPOSED WORK : SVM BASED FAULT DIAGNOSIS OF POWER TRANSFORMER USING OPTIMAL FEATURES

In this section, we first discuss the industrial datasets used by us followed by the architecture of the multilayer SVM classifier to be used to classify the available datasets. We then present our proposed mutual information based feature selection technique before presenting the results.

3.1 Data Available

We have used two industrial datasets for demonstrating the utility of our approach. Each dataset was divided into training and testing data. The training data for dataset 1 was used for selecting optimal features using mutual information. These set of features were then used in dataset 2 as well. For each of the datasets, the SVMs were trained using the training data. Performance was then evaluated using the classification accuracy on the testing data. We use DUPLEX algorithm (Montgomery et al., 2001) to divide a given dataset into training and testing. The DUPLEX algorithm ensures that both the training and the testing subsets contain representative samples from the entire sample space. The available datasets are described next:

Dataset 1: The industrial DGA data for both the normal and fault conditions of power transformers was provided by Crompton Greaves, a power transformer manufacturing company in Mumbai, India. The dataset contains 324 observations of which 214 were selected for training and 110 for testing. Each observation consists of concentration values of seven gases: H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO and CO_2. Table 1 lists this data, labeled as dataset 1.

Dataset 2: This dataset is obtained from Duval and De-Pablo (2001). It contains DGA data corresponding to normal as well as fault conditions from various types of

Table 1. Datasets 1 and 2

Class type	Dataset 1	Dataset 2
	No. of total, train and test samples	No. of total, train and test samples
Normal	[37,24,13]	[25,16,9]
Partial Discharge	[43,28,15]	[26,17,9]
Arcing	[26,17,9]	[45,30,15]
Overheating	[141,94,47]	[10,6,4]
Cellulose Degradation	[77,51,26]	[10,6,4]

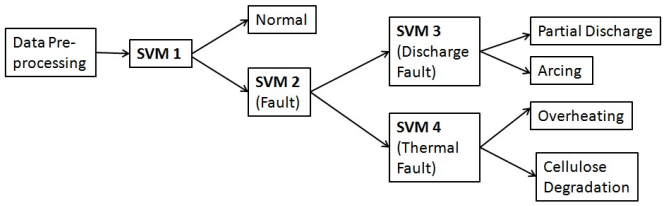

Fig. 2. Architecture of multilayer SVM

transformers installed at different locations around the world. After excluding samples with missing values of any variable, 116 valid samples are obtained of which 75 are selected for training and 41 for testing. Table 1 lists this data, labeled dataset 2. This dataset is not used for feature selection but only to investigate the utility of the optimal features selected using dataset 1.

3.2 Multilayer SVM Classifier

The fault diagnosis of power transformer is a multi-category classification problem. Hence multilayer SVM classifier is used for building a fault diagnosis system for power transformer. The multilayer SVM involves the classification of data into one of the two appropriate classes at each layer (Lv et al., 2005). We propose to develop four SVMs to identify the five classes: normal, partial discharge, arcing, overheating and cellulose degradation. The architecture for the multilayer SVM is shown in Figure 2. SVM 1 separates the normal state from the fault state while SVM 2 separates discharge faults from thermal faults. The third and the fourth SVM classify the discharge faults as either partial discharge or arcing, and thermal faults as either overheating or cellulose degradation, respectively. The preprocessed data (discussed later) is fed to this multilayer SVM classifier for fault diagnosis.

For a given set of features, the following steps are involved in using multilayer SVM classifier for fault diagnosis in power transformers.

1) *Data Preprocessing*: The available data lists the actual value (in concentration units) of the gas content obtained by DGA. For diagnosing faults in a power transformer, the data is processed using Eq (9) to obtain the relative content of gaseous species to be considered instead of the actual values:

$$u_v = \frac{c_v}{\max_{i=1}^{p}(c_i)} \qquad \forall v = 1, 2, \ldots, p \tag{9}$$

where p is total number of variables in the feature set under consideration and c_i denotes the concentration of species i. The rationale behind this scaling is that the fault patterns are related to the relative content of gases rather than their absolute values (Lv et al., 2005). Apart from the given set of scaled features, an additional feature capturing the maximum concentration for a given observation is added. This feature is (Lv et al., 2005):

$$u_{p+1} = log_{10}(\max_{i=1}^{p} c_i) \tag{10}$$

2) *Training the SVMs*: For a specified feature set, the training data after preprocessing (as discussed above) is used to obtain (train) the four SVMs as mentioned in Figure 2. Training each SVM involves finding a hyperplane obtained by solving formulation in Eq 3 corresponding to appropriate classes. For example, to train SVM2, fault data split as either discharge fault or thermal fault is used. The value of regularization parameter C (Eq 3) was chosen to be 100. For obtaining each SVM, Gaussian radial basis Kernel function (described in section 2.1) was used with the value of $\sigma = 1$. These values were taken from Lv et al. (2005). The SVMs are trained using LIBSVM toolbox (Chang and Lin, 2011) in MATLAB®.

3) *Testing the trained SVMs*: The test data is also pre-processed as the training data. Then, the SVM classifier developed based on training data is used to assign the test data to one of the five classes. The performance of the classifier is obtained in terms of the classification accuracy that is the percentage of test data samples that are correctly classified. The classification accuracy is calculated for the complete classifier as well as for the individual SVMs.

3.3 Feature Selection

The performance of the multilayer SVM classifier can be improved by using only the informative variables instead of using all the variables. The informative variables are obtained by using feature selection technique based on mutual information. In case of multilayer SVM, feature selection can be performed in two ways: (i) overall feature selection where the same set of variables is used for all the SVMs, and (ii) pairwise feature selection in which the feature selection technique is applied to individual SVMs and hence the set of variables for each SVM may be different. These two approaches are discussed next. As discussed earlier, the feature selection is performed only for dataset 1.

Overall Feature Selection: It involves the following steps:

(1) Step 1: This step is used for finding the groups with maximum mutual information with specified number of variables in each group. The training data for each of the five classes in dataset 1 is used for calculating the mutual information. The variable selection problem is a combinatorial optimization problem. Since each variable may or may not be selected, there are $2^n - 1$ possibilities for n variables (excluding the case when no variable is selected). For small n, such as in our work, it is possible to explicitly consider all these possible combinations. However, for large n this explicit enumeration is not feasible and heuristic strategies such as greedy search based forward selection (Verron et al., 2008) can be used. In our work, mutual information is computed for all the possible (i.e. $2^7 - 1 = 127$) groups of variables. For each specified value $p = 1, 2, \ldots, 7$ of number of variables, the group with highest mutual information is identified. Thus, after this step, for training dataset

1, we have seven sets of features each corresponding to seven possible values of p. Table 2 lists the resulting seven sets of features.

(2) Step 2: This step is used to compare the classification accuracy for each of the seven groups or sets of features obtained above. For each group the multilayer SVM classifier is obtained using the steps described in section 3.2 and the performance compared based on training dataset 1. These performances are listed in Table 2. The group corresponding to the multilayer SVM giving the maximum classification accuracy is identified to be the optimal set of features.

Table 2. The best group of features with classification accuracies on dataset 1

No. of features	Features used	Classif. accuracy on train, test data (%)
1	C_2H_2	58.87, 34.54
2	C_2H_4, C_2H_2	67.75, 58.18
3	C_2H_6, C_2H_4, C_2H_2	76.16, 59.09
4	CH_4, C_2H_6, C_2H_4, C_2H_2	84.11, 60
5	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2	87.38, 65.45
6	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO	91.12, 73.63
7	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO, CO_2	87.38, 71.81

According to the classification accuracies for training dataset1 listed in Table 2, it is seen that the best multilayer SVM classifier is obtained when the best set of six features is used. This best set of features does not include CO_2, thereby identifying CO_2 as the least informative feature for diagnosing the specified faults.

Table 2 also lists the classification accuracies for test dataset 1 for the listed sets of features. As expected, the performance in each case is inferior to that for the training dataset. It is further seen that the best performance in case of test dataset1 is once again obtained for the set of six features that was identified as the optimal set based on training dataset1. This validates the hypothesis that the performance of the multilayer SVM is improved when the recommended set of six features is used instead of using all the available seven features. .

To further investigate the performance of the classifiers, the performances of each of the individual SVMs used in the multilayer SVM classifier for the training and the testing data for dataset1 is tabulated in Table 3.

The following conclusions can be drawn by analyzing the performance of individual SVMs in this table:
1) The best performance for most of the individual SVMs is obtained for the case when the recommended set of six features are used. For training data of dataset1, this set of six features results in best performance for all SVMs other than SVM4. For SVM4, its performance is marginally inferior to the best set of five features. For testing data of dataset1, the best performance for all the SVMs is obtained with the recommended set of six features. This means that if the user is interested only in distinguishing between a specified pair of faults as considered in the multilayer SVM architecture, the recommended set of six features will again be the best set of features for each SVM.
2) The performance of SVM 1 is very good for the recommended set of six features for both the training (98.13%) and testing data (96.36%) for the dataset 1. Since SVM 1 distinguishes between the normal and fault conditions, it can be concluded that fault detection can be performed with high accuracy with the selected set of features.
3) The performance of SVM 3, which classifies the discharge faults either as partial discharge or as arcing, is 100% most of the times for training data of dataset1. However, for the testing data, the accuracy goes down. This indicates that overfitting is occurring during the training step for this SVM. It might be possible to obtain higher performance for testing data by using a lower value of the regularization parameter C (Eq 3) during the training of the SVM.
4) The performance of SVM 4 is worst amongst all the SVMs for both the training and the testing data of dataset 1. This indicates that the two faults (overheating and cellulose degradation) are not easily separable.

Table 3. Classification accuracy (in %) of individual SVMs on training and testing data for dataset 1

No. of features (same set as in Table 2)	SVM 1 (train, test)	SVM 2 (train, test)	SVM 3 (train, test)	SVM 4 (train, test)
1	88.78, 88.18	85.26, 77.31	100, 66,66	75.17, 52.05
2	88.78, 88.18	86.84, 83.50	93.33, 79,19	87.58, 76.71
3	92.05, 88.18	90.00, 86.59	93.33, 79.19	89.65, 76.71
4	95.32, 90.90	93.64, 85.56	100, 79.19	91.72, 73.97
5	96.26, 92.72	94.7, 90.72	100, 87.50	93.79, 75.34
6	98.13, 96.36	97.36, 91.75	100, 87.50	93.10, 76.71
7	96.26, 95.45	96.84, 91.75	100, 83.30	91.03, 75.34

Pairwise feature selection: We also perform pairwise feature selection for each SVM, where the mutual information is calculated by considering training data for only the corresponding two classes. The set of features giving the maximum mutual information and maximum classification accuracy is selected for that particular SVM. The steps are the same as in the overall feature selection procedure discussed above, with the difference that they are applied individually to each SVM instead of applying them to the overall multilayer SVM classifier. This pairwise procedure results in a different set of features for each SVM containing some or all of the six variables as listed in Table 4. Comparing the results in this table (last row) with the results in Table 2 (second last row), it is seen that similar overall performance is achieved even if different sets of features are used for different SVMs.

Results for dataset 2:
The dataset 2 is now used to verify the performance with the set of six features identified to be optimal based on dataset 1. For this purpose, the multilayer SVM (Figure 2) is trained with the training data of dataset2 when the recommended set of six features are used. The SVM parameters (C, σ) were chosen to be the same as that in dataset 1. Resulting performance for both training and

Table 4. Features obtained for the individual SVMs in multilayer SVM using pairwise feature selection with training data of dataset 1

SVM type	Features used	Classif. accuracy on train, test data (%)
1	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO	98.13, 96.36
2	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO	97.36, 91.75
3	CH_4, C_2H_4, C_2H_2, CO	100, 79.16
4	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2	93.79, 75.34
Overall multilayer classifier	Using above set of features for each SVM	91.58, 71.81

testing data of dataset 2 are listed in Table 5. For the sake of comparison, performances obtained with the best set of five features and all the features are also listed in this table. It is seen that the set of six features leads to acceptable performance, with its performance being the best for training data and slightly inferior to that obtained by the set of five features for testing data.

Table 5. Comparison of the classification accuracies of groups of features for dataset2

No. of features	Features used	Classif. accuracy on train, test data (%)
5	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2	94.66, 70.73
6	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO	98.66, 68.30
7	H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO, CO_2	88, 60.97

4. CONCLUSION

In this paper, a mutual information based optimal set of features is selected for SVM based fault diagnosis of power transformers. It is found that of the seven available gases in DGA, CO_2 should be discarded to obtain better diagnostic performance. This optimal set of six features were identified on an industrial dataset (labeled dataset 1). Diagnostic performance was also analyzed with this set of recommended features on a dataset available in literature (labeled dataset 2) and the performance was found to be satisfactory. The performances for both datasets can be further improved by optimizing the values of some of the parameters associated with SVMs, such as the type of Kernel and its corresponding parameters and the regularization parameter in the SVM formulation. The proposed approach classifies a given observation as belonging to one of the known classes. In case of an occurrence of a new fault, the observation would be wrongly classified to one of the existing classes. The availability of richer data from various classes can avoid such problems and lead to a more robust fault diagnosis scheme. Moreover, based on the results in this paper we expect that while SVMs will have to be trained afresh every time a new dataset is encountered, the optimal set of features as identified in this work can be used. This is of course only for the given types of faults. For other types of faults, feature selection in a similar manner can be performed.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Crompton Greaves, Mumbai for supporting this work.

REFERENCES

Bacha, K., Souahlia, S., and Gossa, M. (2012). Power transformer fault diagnosis based on dissolved gas analysis by support vector machine. *Electric Power Systems Research*, 83(1), 73–79.

Chang, C.C. and Lin, C.J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Duval, M. (1989). Dissolved gas analysis: It can save your transformer. *IEEE Electrical Insulation Magazine*, 5(6), 22–27.

Duval, M. and DePablo, A. (2001). Interpretation of gas-in-oil analysis using new iec publication 60599 and iec tc 10 databases. *IEEE Electrical Insulation Magazine*, 17(2), 31–41.

Erdal, K., Okan, O., Omer, U., and Thomas, D. (2009). PCA based protection algorithm for transformer internal faults. *Turkish Journal of Electrical Engineering and Computer Sciences*, 17(2), 125–142.

FIST (2000). Transformer maintenance. *Facilities Instructions, Standards, and Techniques, Hydroelectric Research And Technical Services Group , Bureau of Reclamation, Denver, Colorado*, 3-30.

Han, H., Wang, H., and Dong, X. (2011). Transformer fault dignosis based on feature selection and parameter optimization. *Energy Procedia*, 12, 662–668.

IEEEGuide (2009). IEEE guide for the interpretation of gases generated in oil-immersed transformers. *IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991)*, C1 –27.

Lv, G., Cheng, H., Zhai, H., and Dong, L. (2005). Fault diagnosis of power transformer based on multi-layer svm classifier. *Electric Power Systems Research*, 75(1), 9–15.

Montgomery, D., Peck, E., and Vining, C. (2001). *Introduction to Linear Regression Analysis*. John Wiley and Sons, third edition.

Naresh, R., Sharma, V., and Vashisth, M. (2008). An integrated neural fuzzy approach for fault diagnosis of transformers. *IEEE Transactions on Power Delivery*, 23(4), 2017–2024.

Rogers, R. (1978). IEEE and IEC codes to interpret incipient faults in transformers, using gas in oil analysis. *IEEE Transactions on Electrical Insulation*, 13(5), 349–354.

Vapnik, V.N. (1995). *The nature of Statistical Learning Theory*. Springer Verlag, New York.

Verron, S., Tiplica, T., and Kobi, A. (2008). Fault detection and identification with a new feature selection based on mutual information. *Journal of Process Control*, 18(5), 479 – 490.

Zhengwei, Z., Zhenghua, M., Zhenghong, W., and Jianming, J. (2009). Model study of transformer fault diagnosis based on principal component analysis and neural network. In *International Conference on Networking, Sensing and Control, Okayama, Japan, 2009*, 936 –940.