A Novel Feature Selection Method for Fault Diagnosis

Zacharias Voulgaris[1], Chris Sconyers[1]

[1] ICSL lab, Georgia Institute of Technology, 813 Ferst Drive NW,
Atlanta, GA 30332, USA.
{zvoulgaris, csconyers}@gatech.edu

Abstract. A new method for automated feature selection is introduced. The application domain of this technique is fault diagnosis, where robust features are needed for modeling the wear level and therefore diagnosing it accurately. A robust feature in this field is one that exhibits a strong correlation with the wear level. The proposed method aims at selecting such robust features, while at the same time ascertain that they are as weakly correlated to each other as possible. The results of this technique on the extracted features for a real-world problem appear to be promising. It is possible to make use of the proposed technique for other feature selection applications, with minor adjustments to the original algorithm.

Keywords: feature selection, dimensionality reduction, fault diagnosis, classification

1 Introduction

The extraction of features is a relatively manageable and straightforward process, especially in cases where there is abundance of data. Also, as there have been developed methods for automating this process, the feature set size increases dramatically (leading to what is commonly refer to as the *curse of dimensionality*). However, the often large number of features does not necessarily mean better results in one's analysis of the problem at hand. Many of the extracted features are usually redundant and even relatively weak. Since the fault diagnosis process, in order to be carried out effectively requires one or more robust features; it is essential that an efficient feature selection step is required and demands the researcher's attention. Fault diagnosis based on the selected feature is usually carried out after the stage of feature fusion, which involves taking the selected features and creating a super-feature by combining them. However, in this paper the selected features are used as they are, so that they can be evaluated through their diagnosis classification performance.

Feature selection has been tackled mainly as an optimization problem. The relevant research includes supervised, unsupervised and semi-supervised ways to select the optimum subset of features [1]. Yet, the general approaches to feature selection are twofold: either rank the features according to some evaluation metric and select the top k ones, or select a minimum subset of features that allows for good performance (i.e. the quality of the dataset, in terms of relevant information, does not deteriorate). [1]. Furthermore, the feature selection techniques, particularly in the fault diagnosis/prognosis domain, can be categorized in two broad groups: optimal and sub-optimal methods, the former including exhausting search and the Brand & Bound algorithm, while the latter using methods like sequential search, floating search, plus-L-minus-R search [2].

The rest of the paper is structured as follows. In section 2 the related work to feature selection methods is presented and briefly analyzed, after the evaluation criteria for the features at hand are defined. Based on the analysis of the feature selection methods, the necessity to adopt an alternative approach is pinpointed for the application at hand. This approach is described, on a theoretical level, in section 3. In the section that follows (4), the application of the proposed method is exhibited, using some real world data from a project the authors are involved in. The results of the relevant simulations are also presented in this section. Afterwards, a discussion of this research is carried out in section 5. Conclusions are drawn and avenues of future research based on this work are mentioned in the final section of this paper (6).

2 Evaluation Criteria and Related Work

There have been numerous attempts to solve the feature selection problem using various strategies for obtaining a robust reduced feature set. Although most of these approach are classification-oriented (including data mining) [3-5], some are designed having a fault diagnosis/prognosis problem in mind [2, 7-9].

Although the basic principles are the same in these disciplines, as regards the feature selection problem, in the later case the evaluation of a feature is more complicated. Namely, a robust feature in fault diagnosis (and prognosis) is characterized not merely by good distinguishability (discernibility, class separability, etc.) but by a sense of monotonicity as well (the more monotonic in relation to wear level the better). As monotonicity can be modeled by taking the relationship between consecutive wear levels (classes) so that the feature values form a monotonic pattern, it can be thought of as *ordered* discernibility. So far, the most widely-used and academically acceptable measure of monotonicity is the correlation coefficient.

There are several methods for dimensionality reduction, which are not related to the problem at hand (e.g. PCA yielding a small number of features to express almost the same information through different combinations of the original features). The term feature selection which is used in this research denotes that a subset of the *original* features is used, thereby maintain many of the original dataset characteristics and yielding more meaningful features in the new feature set.

Classification-oriented methods for feature selection include, but are not limited to, vertical compactness of data, the relevance-based algorithm approach, statistics-based techniques, and Genetic Algorithms [3]. Alternative approaches in the same category of methods for feature selection include the Receiver Operating Characteristics curve approach and class separability measures, through certain searching techniques (which are usually suboptimal) [4]. As regards class separability measures, one of the authors of this paper has conducted a thorough research on one of them (which he coined with the term Discernibility) with applications in feature selection, among a number of other research topics in the field of classification [5].

Feature selection for fault diagnosis/prognosis [6], though exhibiting similarity in the methods it employs, often makes use of alternative ways to tackle this problem. Namely, in [2] an enhanced version of the Brand & Bound algorithm is proposed and implemented. Also, the use of consistency as a measure of feature evaluation for feature selection has been successfully employed, as it exhibits monotonicity and speed [2]. A Genetic Programming based approach has also been applied for feature selection, as it yields good results without the need for large samples, nor assumptions used in e.g. statistical approaches to the problem at hand [7]. Also, this approach is preferred by some researches due to its efficiency since it bypasses the analytical way of creating all possible feature subsets, which is an NP-hard problem [7]. Finally, a rough sets approach has also been employed for feature selection in this domain, with promising results [8].

It should be noted that feature selection is application dependent, especially in fault diagnosis/prognosis problems [9]. Yet, there are some recurring patterns that are found across the different applications in the field. These are the basis for feature analysis and evaluation, and comprise of the following criteria [9]: Distinguishability (often referred to as class separability or Discernibility), Detectability/isolatability, Identifiability, Degree of Certainty. The use of at least one of these criteria is essential for conducting a proper feature selection for the field of fault diagnosis/prognosis.

Note that most of the aforementioned methods for feature selection assume quantitative feature values. This is the most commonly encountered feature type in applications related to fault diagnosis. However, there exist several other feature types, which may require a different strategy for feature selection [10].

Though all of the above methods show promising results in the applications they were tested, a generic and therefore versatile method of feature selection is yet to be found. The fact that feature selection often requires some type of threshold to filter out the redundant features is part of the problem. Another aspect of the problem is that the feature selection methods under consideration are based on one or, less often, more evaluation metrics that model the four criteria mentioned earlier. If the user of these methods wishes to experiment with alternative measures for ranking the features, this would be a time-consuming task which in some cases it is even infeasible, due to the nature of the feature selection method. An attempt to address these issues is made using an alternative, more generic approach to feature selection, which is presented in this paper.

3 Methodology

The proposed method performs feature selection in an efficient and automated way (which is why it is called Automated Feature Selector or AFS for short). AFS takes as inputs the feature matrix *F* (input values in classification terminology), the wear level vector *W* (target values) and, optionally, a correlation threshold *th*. The later is there in case the user wishes to obtain only features that meet a certain quality standard (e.g. strong correlation with the wear) and can be particularly useful in cases of very large feature sets.

The AFS method functions as follows. Initially, a dimensionality check is performed, to ensure that both *F* and *W* contain the same number of points (*N*). Once this is attained, the features are individually evaluated to determine how useful they are for the objective at hand. This is done by first calculating the correlation between each feature *i* (F_i) and *W* and storing it in a vector *C*. Then, the standardized Fischer Discriminant Ratio of *Fi* is calculated as well and stored in vector *DR*. By the term standardized we mean a variation of FDR such that it yields values in the interval (0, 1], for reasons that will become apparent later on.

Following this stage, a series of similarity matrices for all feature combinations is calculated, in order to determine how related these features are in terms of the information they yield. This step comprises of the computation of three distinct metrics (in the form of three matrices). The metrics used are absolute correlation, similarity (with $\alpha = 1$) [6], and 1 – cosine of angle they form [10]. These metrics were selected so that a significant amount of information about the features can be extracted. Yet, even though this information is more than the information yielded by a single metric, it can be increased if more metrics are added to the list, as long as they all yield values in the interval [0, 1].

After this, a likeliness metric *L* is calculated by taking the point-to-point product of these matrices and adding I_n to it, where *n* is the number of features in the original feature set. A (feature) potency metric *G* is also computed, by taking the point-to-point product of vectors *C* and *DR*. By taking the ratio of the aforementioned two composite metrics, *L* and *G*, yields a distance-like metric *D*, which constitutes the basis of the feature selection process that ensues. This metric is in essence the evaluation of all the individual features of the original feature set and its presence reduces the original problem into a relatively simple mathematical problem, which is solved in the steps that follow.

Afterwards, the most correlated feature *f* is selected and its index is stored in vector *in* (which is the index vector, one of the outputs of the AFS method). This is done because in fault prediction correlation with the wear level plays a vital role for assessing the quality of the condition indicator at hand, so as a starting point one would prefer to have something quite robust. Alternatively, the feature with the highest potency could be selected instead. The inverse of the correlation value of this feature with the wear is stored as the first element of vector *X* and all “distances” from *f* to the other features are set to infinite. This is done so that feature *f* is not selected again. Following that, the “closest” feature to *f* (let us call it *g*) is found and its index is stored in *in*. The “distance” between *g* and *f*, in terms of the distance-like metric D, is stored in variable *m*. The average of X_1 and *m* is stored in X_2 and all “distances” to *g* are set to infinite (so that feature *g* is not selected again either). By repeating the above process for all *n* features a distance-based feature structure is formed in matrix *X* and vector *in* contains the corresponding indexes.

If a correlation threshold *th* is given, the features from *in* that yield a *C* value greater or equal to *th* are selected while the rest of the features are discarded. Otherwise, the smallest value of *X* is found (as well as the corresponding index *ind*) and all features having an *in* value between 1 and *ind* are selected. In essence this means that the feature set is partitioned in two clusters, based on *D* instead of their spatial distances. One of these clusters contains the “good” features, which are taken as an output of the method, while the other contains the “bad” ones, which are discarded. Alternatively, other methods of clustering could be used to achieve this binary partitioning, but this one was selected for its simplicity and effectiveness in practice.

4 Experiments, Results and Discussion

To test the effectiveness of this method, a set of features extracted from a real-world dataset was used. The features were based on vibration data of axial and radial orientation of the sensors in relation with the bearing examined. The wear levels, four in total including the baseline, corresponded to different corrosion levels of the bearing. Also, the data included two distinct working conditions and were collected over a number of experiment rounds. The feature sets used in this research comprised of 16 features for radial and 16 features for axial data. In both cases they comprised of 240 data points (larger dataset could have been tested by increasing the sampling rate of the features during the extraction process, but this would not change anything in the whole feature selection process, due to the nature of the features in this field). From these feature sets two feature selections were made, one for each set. In the first selection (SF_1) which corresponds to the axial features, six features were selected by the AFS method, while for the radial features, the new feature set (SF_2) comprised of four features. An additional feature selection was carried out afterwards, this time making use of the union of the original two feature sets. This yielded a subset (SF_3) made up of four features (two from each one of the original feature sets).

The experimental setup for the feature selection part is as follows. Using the 240 data points of the dataset, we split them into 10 equal subgroups, using k-fold cross validation, allowing for 216 training points and 24 test points in 10 cross validation groups. These groups are all separately classified across all features within the feature set through a number of classifiers to compare classification performance.

Four classifiers were chosen based on relative speed and accuracy of classification—three of them different flavors of the *k*-Nearest Neighbor algorithm, thoroughly described in [5]. For the variants that made use of a k parameter (number of neighbors), the value of 5 was chosen, since it is one commonly used. The classifiers used are the following:

- C4.5 Decision Tree Classifier — (C45)
- Variable *k*-Nearest Neighbor — (VKNN)
- Weighted *k* -Nearest Neighbor — (WKNN)
- Discernibility *k* -Nearest Neighbor — (DKNN)

All six feature sets are classified separately: reduced feature sets for radial, axial, and both features, and total feature sets for the same. For each classification is computed the classification accuracy, measured as percent correct test point classifications; and the CPU time, measured as the number of seconds per classification round. These will measure the ability of the reduced feature set and classifier to properly diagnose the wear level, as well as how long a classifier requires to train itself for diagnosis. These are computed and averaged across 30 iterations of the above procedure, resulting in a total of 300 classification experiments for each combination of feature set and classifier.

A brief analysis of the features selected from the above subgroups of the original feature set is shown in Table 1.

Table 1. Feature Characteristics. Discernibility refers to the Spherical Index of Discernibility [5], a measure of class distinguishability.

Feature	Correlation with Wear	Fischer Discriminant Ratio	Discernibility (SID)	Comments
# 3, Axial	0.8542	0.9880	0.7292	SF_1, SF_3
# 4, Axial	0.6242	0.8286	0.2833	SF_1, SF_3
# 9, Axial	0.6859	0.8198	0.3667	SF_1
# 12, Axial	0.6468	0.7211	0.2375	SF_1
# 13, Axial	0.8194	0.9733	0.5833	SF_1
# 16, Axial	0.6759	0.7787	0.4792	SF_1
Mean, Axial	*0.6853*	*0.7996*	*0.3352*	*All features*
# 3, Radial	0.7460	0.9662	0.2583	SF_2
# 5, Radial	0.6400	0.9992	0.5375	SF_2
# 8, Radial	0.7082	0.7930	0.2583	SF_2
# 14, Radial	0.7229	0.9016	0.1833	SF_2, SF_3
# 15, Radial	0.6312	0.9993	0.5458	SF_3
Mean, Radial	*0.6871*	*0.8514*	*0.3003*	*All features*

Further analysis in the feature subsets yielded by AFS revealed some interesting information that provides insight to the collaborative potential of the selected features. Specifically, the Spherical Index of Discernibility (SID) [5] was employed, first on the whole feature sets and then on the subset of selected features. As it is expectable, there was a drop in most cases in the SID value when the feature set was reduced, yet this drop was not dramatic, considering the dimensionality reduction ratio (Table 2). Yet, in the case of the combined feature set of 32 features, the reduced feature set yielded a surprising increase in the SID, despite the significant dimensionality reduction ratio exhibited (Table 2).

Table 2. Results of Classification Experiments and Feature Set Evaluation. The Index of Discernibility refers to the Spherical one, introduced and described in [5]. The numbers in bold denote an improvement.

Feature Set	Mean Accuracy	CPU time	SID	Dimensionality Reduction Ratio
Axial	94.3%	0.530s	0.6625	-
Radial	72.0%	0.534s	0.2792	-
Both	85.6%	0.984s	0.2958	-
SF_1 (Axial)	85.5%	**0.222s**	0.5000	62.5%
SF_2 (Radial)	69.9%	**0.152s**	0.2583	75.0%
SF_3 (Both)	**86.7%**	**0.160s**	**0.6208**	87.5%

Figure 1 shows the accuracy rate as a percentage of the number of test cases classified correctly per reduced feature set and for each classifier. The feature sets for radial features and for both features, when classified with a KNN-based classifier, show around 90% accuracy in diagnosing the correct wear level.

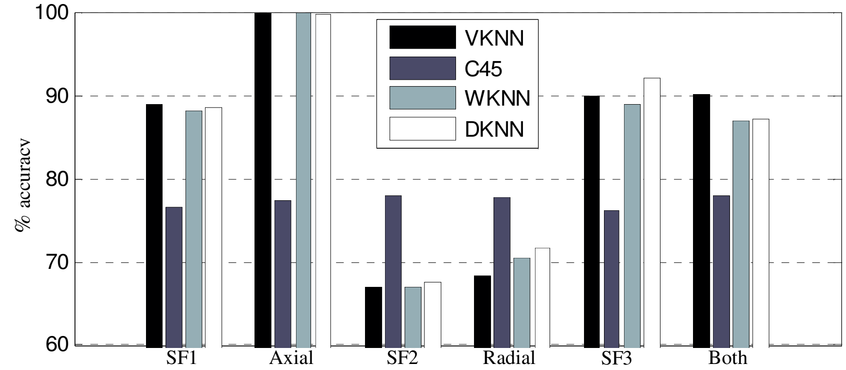

Figure 1: Accuracy of classifiers on reduced and complete feature sets.

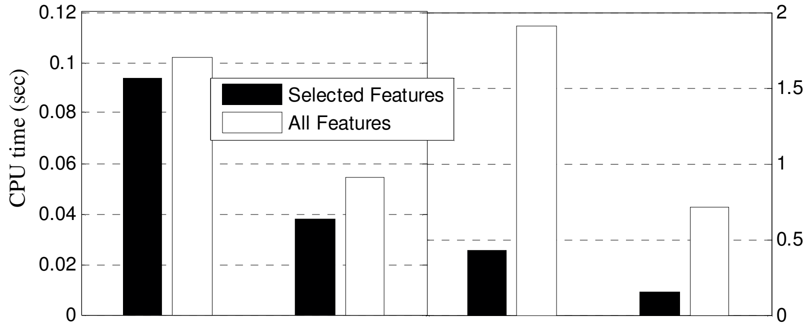

Figure 2: Average classification CPU time. Classifiers are, from left to right: Variable kNN, Discernibility kNN, C45, and Weighted kNN.

The time it takes each classifier to train on each feature, as a measure of processing time in seconds, is computed and averaged for all reduced feature sets, and all complete feature sets. The CPU time (Figure 2) shows definite improvement in training time for all four classifiers. Particularly, The C45 and Weighted kNN classifiers, which exhibit an almost 4 times speed up in training time.

6. Conclusions and Future Work

Based on the previous analysis, it can be concluded that the proposed method yields robust feature subsets, something that reflects in good accuracy rates for fault diagnosis, using a number of classifiers. As one would expect, the accuracy rates in the reduced feature sets are not as high as in the original feature set, since there is some information loss in the feature selection process. However, in the case of the complete feature set of 32 features not only there is no decrease in the accuracy but the opposite is observed. It is also noteworthy that the reduced feature sets are much smaller than the original ones, something that yields significantly better classification speed (based on CPU time). Moreover, the Discernibility of the reduced feature sets appears to not drop significantly, while in one case even increase, after the feature selection takes place.

Future work will include alternative measures of (dis)similarity to incorporate in the AFS method. Also, further testing of this method, in different fault feature sets will be conducted to test its generality. Furthermore, it is planned to make use of AFS as a module of a larger system, which will be implemented as an autonomous, fully automated feature generation system taking as inputs preprocessed data and yielding a single fused feature, along with its characteristics, as outputs.

Acknowledgement

The authors of this paper would like to acknowledge the sponsor of the AVDPIP project, from which the feature data used in this research stems. Also, they would like to thank the rest of the research team involved in AVDPIP, in both the Georgia Institute of Technology and in Impact Technologies.

References

1. Liu H. and Motoda H.: Computational Methods of Feature Selection, Chapman & Hall/CRC. Boca Raton (2008)
2. Liu, X.: Machinery Fault Diagnostics Based on Fuzzy Measure and Fuzzy Integral Data Fusion Techniques. PhD thesis, Queensland University of Technology (2007)
3. Liu, H., Motoda, H.: (editors): Feature Extraction, Construction and Selection – A Data Mining Perspective. Kluwer Academic Publishers, USA (1998)
4. Theodoridis S., Koutroumbas K.: Pattern Recognition. Academic Press, USA (1990)
5. Voulgaris, Z. N.: Discernibility Concept in Classification Problems. PhD thesis, University of London (2009)
6. Vachtsevanos, G., Lewis, F. L., Roemer, M., Hess, A., Wu, B.: Intelligent Fault Diagnosis and Prognosis for Engineering Systems. John Wiley & Sons, Inc, Hoboken, NJ (2006)
7. Sun, R., Tsung, F., Qu, L.: Combining Bootstrap and Genetic Programming for Feature Discovery in Diesel Engine Diagnosis. International Journal of Industrial Engineering, vol. 11(3), pp. 273-281 (2004)
8. Lee, S.: An Architecture for a Diagnostic/Prognostic System with Rough Set Feature Selection and Diagnostic Decision Fusion Capabilities. PhD thesis, Georgia Institute of Technology (2002)
9. Saxena, A., Vachtsevanos, G.: Optimum Feature Selection and Extraction for Fault Diagnosis and Prognosis. Proceedings of the 2007 AAAI Fall Symposium on Artificial Intelligence for Prognostics, Arlington, VA. (2007)
10. Pekalska, E., Duin, R. P. W.: The Dissimilarity Representation for Pattern Recognition Foundations and Applications. World Scientific, Singapore (2005)