2011 11th IEEE International Conference on Data Mining Workshops

MODELING UNRELIABLE DATA AND SENSORS: Using F-measure attribute performance with test samples from low-cost sensors

Vasanth Iyer* and S. Sitharama Iyengar
*International Institute of Information Technology, Hyderabad, India - 500 032
Louisiana State University, Baton Rouge, LA 70803, USA

vasanth@research.iiit.ac.in,iyengar@csc.lsu.edu

***Abstract*—Building a high performance classifier requires training with labeled data, which is supervised and allows generalizing the classifier's decision boundary and in practice most of the data is unlabeled, newer algorithms needs to be learn by knowledge discovery. Sufficient training data are collected in the form of empirical evidence, which have labeled positive and negative samples to build the hypothesis. The hypothesis is constructed by the conjunction of the attributes, which can be learnt by machine learning algorithm. In this paper, we work with two forms of ranking weights, precision and relevance, which help in finding hidden patterns and prediction future events. Empirical evidence for a weather patterns and tracking of a phenomenon needs to accurately extract the attributes and label the training samples, which is a very laborious and time-consuming effort. Automating weather prediction algorithms, which are trained by supervised learning, needs to be generalized so that it can be tested with unreliable and noisy weather data from low-cost sensors. We use a training data from previous forest fires events, the datasets containing all the attributes are labeled using manual data logs for a given geographical area. The labeled original dataset is mapped to the data collected from on-line sensors, which further improves the accuracy of the training set. As some of classes have very few samples, which are related to the peak fire seasons, domain specific knowledge are added by sensor measurements and Fire Weather Index (FWI) to help accurately model the events. We show that training accuracy of the small forest fire classifier using attributes from manual logs is enhanced by 30% by using sensor data. The rare and hard to classify large forest fires are 95% accurately classified by using the new Fire Weather Index (FWI). We also show that our framework is more robust to outliers from noisy sensor measurements by accounting for in the model parameters. The model allows further generalization for linearly and non-linearly separable datasets by estimating the parameters $(1-\delta)$ and minimum allowable error ϵ for hypothesis, sampling accuracy and cross validation.**

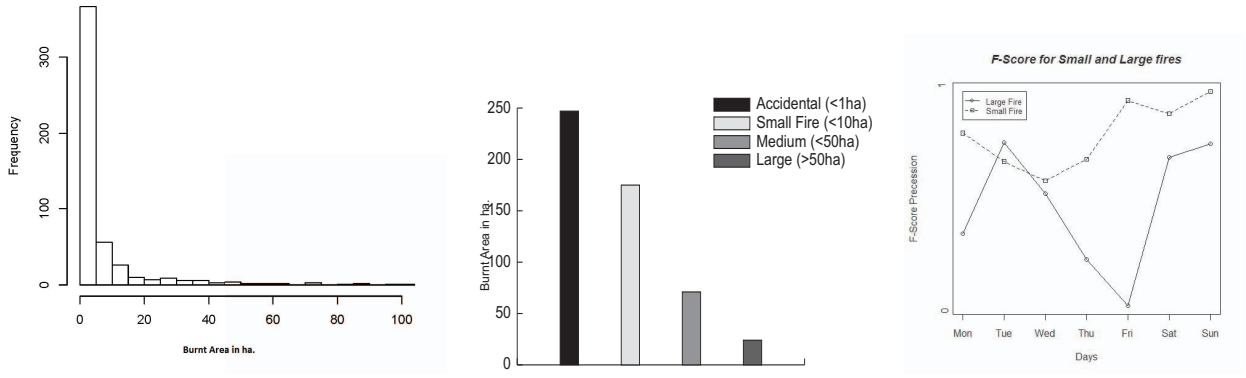
***Index Terms*—Sampling sensors, Data mining, Machine Learning, Ranking functions, Knowledge discovery, Event Modeling, Forest fires, FWI, Temporal Patterns, Sensor Networks.**

I. INTRODUCTION

Data obtained from weather stations have high dimensionally and are very noisy, often designed with remote sensing as the primary objective. These systems are not suitable for accurate monitoring and tracking of the weather data and a detail modeling is necessary to adapt such systems to help track and predict future events accurately. To express the features of the original problem in a linear space, the dataset is transformed into a vector into a higher dimension, which helps to separate different categories and learn the coefficients of the classifier's decision boundary. These coefficients are then applied to the original dataset dimension space with the estimated model bias and variance to predict new test cases. Algorithms like linear discriminant analysis (LDA) and Principal Component Analysis (PCA) have been used to pre-process large datasets to efficiently select only relevant attributes, which help better represent the model avoiding data redundancies. This allows the designed algorithms to perform well as the complexity of computation is reduced, in a similar manner we use statistical techniques to sample the training data so that only fewer features are sufficient to reproduce the variation in the input signal. Weather model assumes the temporal features sets are highly correlated and does not follow the normal assumptions of independently identically distributed (i.i.d) sampling. Combination of attributes to learn the concept hypothesis without over-fitting the data helps better performance during testing.

The performance of the classifier can be evaluated in terms of how well the algorithm generalizes the decision boundary for a given training set data, so it performs well during the testing. In our case the testing samples are generated by low-cost sensors with a given accuracy, which is determined by the number of sensors and the noise level. If the accuracy is defined as $(1-\delta)$ and the minimum error allowed ϵ, then a cost function matrix [8] can be used in the classification to further model the false alarm rates. The model uses temporal features and correlated sensor measurements to better model the target function of fire activity. The cause of high error rates are due to overlap in class densities in the original feature space [12], transforming to a higher dimension feature space allows better approximation in such cases. As the model assumes linear separability of the dataset, due to errors the weights are no longer linearly dependent, appropriate polynomials kernels are used to defined the decision boundary. The accuracy performance bench marking is carried out for different classifiers and the scores are compared with kappa score and misclassification error rates. To further, investigate

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.124



(a) Empirical log collection. **(b) Four class classification.** **(c) F-score for temporal features.**

Fig. 1. Histograms of empirical fire activity and its labeled class categories showing F-score.

dependent factors that can cause such fires, we include statistics, which are the number of visitors and traffic patterns coming into the forest area. This adds to the model the knowledge of temporal activity, measurable temporal patterns and operating thresholds useful in fighting fire events.

The rest of the paper is organized as section II and III provides background and related work. Section III introduces ranking of the original dataset with the help of machine learning and section IV introduces Data Mining and the use of inexpensive sensors to increase the reliability of classification of the original dataset. Section V uses an extended Fire Weather Index dataset, when the dataset overlaps and is not separable for rate events. Section VI and VII introduces linear and kernel functions to address generalization errors and cost based classification and discusses the results using WEKA framework. The paper concludes with a brief summary followed by acknowledgements in sections VIII and IX.

II. Background

Forest fire event detection is of primary importance to unsupervised data collected from sensor network deployment in remote areas. The original dataset we use to conduct experiments are taken from UCI Machine learning repository [4], which has the number of past fire events since 1994-2001 for a given park area. The log contains the fire activity and the number of hectors of the burn area during the year, selected samples that best represent the population and it's geographically relevant attributes including its frequency is then recorded to help classify the type of fire.

III. Event ranking using machine learning

The current event data log used to for forest fires may be incomplete and how does one know the distribution knowledge and patterns of events from the data. Relevance factors of fire event concept can be defined for all fire events that are tagged in the log, as we are interested in ranking fire occurrences in the record. The inverse concept precision is calculated from the histogram plot shown in Figure 1(a), plot shows there are very few large fires. When reporting on major fire events the most relevant samples are retrieved, which has a higher rank leading to a precision score [12,13,20] close to 1.

A. Linear ranking function

To design a good ranking function it needs to balance the relevance and precision of the events in a way to express a summable numerical quantity, which is statistically perfect. The new weights are evaluated for each fire types and the records, which match the user query [9] is sorted and returned.

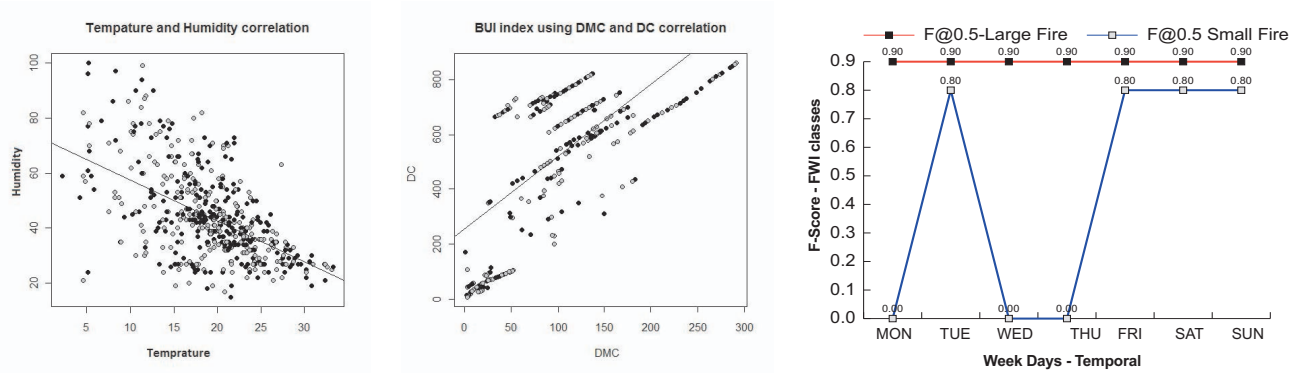
$$\text{Precision} = \frac{\textbf{number of relevant forest fire events retrieved}}{\textbf{number of forest fires retrieved in query}}$$

$$\text{Relevance} = \frac{\textbf{number of relevant forest fire events retrieved}}{\textbf{number of relevant forest fires classified}}$$

B. F-score based feature performance

The weighted harmonic mean of precision and relevance are used to compare the performance of two class labels, the traditional F-measure or balanced F-score is because recall and precision are evenly weighted. Figure 1(c) and Figure 2(c) shows F-score variations with temporal week day attribute. Higher the score the better the classification accuracy, the sensor dataset shows small fires are better learnt and the FWI dataset shows the large fires reliably detectable. The general formula for non-negative real β is:

$$\text{F}_\beta = \frac{2 \times \ Precision \times Recall}{Precision + Recall} \tag{1}$$



(a) Negatively correlated. **(b) Positively correlated.** **(c) Performance of large fire.**

Fig. 2. Feature correlation for sensor and FWI datasets and large fires are shown with the new FWI F-score.

It is based on van Rijsbergen's [20] effectiveness E given by

$$\text{Alarm}_{\text{rank}} = 1 - \frac{1}{\alpha\frac{1}{p} + (1-\alpha)\frac{1}{r}}$$

We further look into accidental small fires, as they are very probabilistic and any conceptual link to the attributes may lead to rank the idea of precision and relevance of the collection. $\alpha-$ the alarm weight are calculated based on the reliability of the ground truth, higher precision weights is given due to prior assumption that large and medium fires are not accidental but correlated to natural weather patterns. The precision weight factor for small fires can then be evaluated given α by $(1-\alpha)$ for a given burnt area(BA). A continuous valued function for BA can be defined as

$$BA_{Days\%} < \textbf{50 ha} = \theta_1 + \theta_2 \ln(NF) \tag{2}$$

where $\theta_1 = 2.895$ and $\theta_2 = 1.265$, which is the variance of the BA data versus fire activity, showing logarithmic $O(\lg(BA)$ complexity as shown in Figure 1(a).
We a learning the two major categories fire categories, which are accidental small fires and large wild fires. To select the performance of the two we use weighted precision versus relevance to estimate the ranking information F@(0.5) for the above equation (1). The F-scores are calculated and weighted for high reliability by using F@(0.5), which is twice the precision compared to its equivalent relevance scale. In equation (1) reliability and precision are proportionally weighted, while relevance is inversely proportional. The performance scores shows in Figure 1(c) for accidental small fires is 3 times higher when compared to queries for large wild fires or from the same collection.

C. Performance based on user selectable temporal attributes

Accidental small wild fires are possible all through the year, making is a viable application for automated sensor measurements. The measurements such as temperature, humidity and wind gust are automated, while temporal attributes such as human traffic, day of the week are used to study the small fire patterns. The ranking suggest that small fires events are mostly due to temporal causes such as human traffic and vehicular routes more than any observed dataset or correlated sensor measurements.

IV. Data mining of sensor data

The original dataset described represents the ground truth, due to unknown data distribution estimating mean μ, of the original population will lead to errors as discussed in Definition 3. The initial results shows that the cause of fire activity is related to temporal activity, which is hard to measure. To explore related features of weather and climate data we use previously collected data from automated logs of sensor network [1,5]. The attributes which are of interest are temperature, humidity, rain and wind conditions during the time of the fire event. The sensors attributed are measured values and easily available due to existing infrastructure of weather stations. This data mining approach allows to add quantitative knowledge to the original dataset hypothesis, such as a GPS position information to a weather dataset. To learn the new attributes of the fire event hypothesis is not very practical in an experimental setup as sensor data is prone to noise. The measured attributes are highly correlated and do not follow the normal independently identically distributed (i.i.d) assumptions of sampling, as described in Definition 1. To measure correlated sensor measurements of interest, interval estimates from multiple reading are necessary and overlap with the i.i.d of other sensors to calculated the overlap precise interval and range accuracy, as described in Definition 2. In an event of a fire event the two parameters temperature and humidly is shown in Figure 2(a) may not match the expected mean (μ) and variance (σ). This further does not help the classification accuracy as both the parameters are uniformly distributed over all the classes having mean $\mu_{temprature} = 30$. Due to

	Sensor Model Parameters				Dataset Accuracy						
Experiments	**Features**				**Training Error** $\delta_{train}\epsilon_{train}$						**True Error** ϵ_{sample}
	Class	Attr.	Performance	SV	AF	SF	MF	LF	err	kappa	All classes
Hypothesis	4	3	F-measure	517	17%	17%	0.9%	0.9%	-	-	-
Hypothesis	2	3	Subset	422	62%	62%	-	-	37%	0.14	56%
DM (NB)	2	6	Sensor	517	84%	30%	7%	0%	48%	0.13	47%
Weka(J48)	4	9	Sensor	517	92%	67%	40%	29%	28%	0.53	47%
Weka(SVM)	4	9	Sensor	480	97%	93%	94%	96%	3%	0.95	45%
Weka(SVM)	4	9	Poly Kernel	411	95%	93%	94%	95%	5%	0.90	49%

Class=Multi-class Attr=No. of attributes SV=Support Vectors AF=Accidental Fire SF=Small Fire MF=Medium Fire LF=Large Fire err=Misclassification kappa=score

TABLE I
F-SCORE PERFORMANCE WHEN USING SENSOR DATASET.

this shortcoming fire activity related parameters are also needed to predict, the FWI dataset discusses these model issues, which has unique class μ.

Measurements 1 *The sensor network model measurement matrix maintains $i.i.ds$, as there are lots of correlated readings a sparse model is used for a collection of sensor. The collection $x = P\theta$ is for all possible basis representation, which can be measured with an allowable error is called the sparsity model.*

Measurements 2 *The sparsity measurement matrix maintains non-overlapping subsets, which are present in all signals and among all cost (bits) level representation $X = P\Theta$ of the processed signal. The non-overlapping coefficients represents the basis, which is the lossless representation of the measured signal ensemble.*

Definition 1 *Interval estimate-Computation: For a given ensemble X, we let $\mathrm{P}_F(X) \subseteq \mathrm{P}$ denotes the set of feasible location matrices $P \in \mathrm{P}$ for which a factorization $X = P\Theta$ exits. We define the joint sparsity levels of the signal ensemble as follows. The joint sparsity, level D of the signal ensemble X is the number of columns of the smallest matrix $P \in \mathrm{P}$. In these models each signal x_j is generated as a combination of two components: (i) a common component z_C, which is present in all signals, and (ii) an innovation component z_j, which is unique to each signal. These combine additively, giving $x_j = z_C + z_j, j \in \forall$. $X = P\Theta$. A further optimization can be performed to reduce the number of measurement made by each sensor, the number of measurement is now proportional to the maximal overlap of the inter sensor ranges and not a constant. This are similar to training accuracy and errors of the original weather samples obtained by local stations. This is calculated by the common coefficients K_c and K_j, if there are common coefficients in K_j then one of the K_c coefficient is removed and the common Z_c is added, these change does not affect the reconstruction of the original measurement signal x.*

A. Correlated features

The original samples of the fire events are sorted in terms of BA in hectors(ha), the frequency distribution versus BA is shown in Figure 1 and its parametric model is given earlier in equation (2). The parameters θ_1, θ_2 are found using linear regression techniques it is still difficult to build and relate to weather model. We can model the BA distribution in equation (3) in terms of a dependent function $f(x)$, which uses weather attributes as given in equation (4) and equation (5). This is the first integration of the weather model to the prior fire activity, this model not only classifies the training set but allows to continuously monitor and predict fire activity using inexpensive sensors over wireless networks.

$$BA = f(x) \tag{3}$$

B. Labeling BA in hectors.

The histogram shows that the BA is skewed with large number of small fires and very few large fires, predicting likelihood of small fires learnable. We further classify fires into four categories, which allows to compare the performance of the system in terms of precision and relevance. The Figure 1(b) shows the new histogram of the BA in terms of four class labels, which are used in the prediction of new weather samples obtained from the local sensors. The accuracy of classifying fire events in the training set with limited feature set is only 52%.

C. Estimating BA with temporal and correlated attributes

The total number of independent variables in the fire activity target function is seven as given in equation (6).

$$\mathrm{V}_{\mathrm{train}}(\mathrm{FireEvent}_{\mathrm{Day\ of\ week}}) \leftarrow \hat{V}(FireEvent_{Day\,of\,week})\textit{Temporal Variable} + \hat{V}(FireEvent_{Day\,of\,week})\textit{Correlated measurements} \tag{4}$$

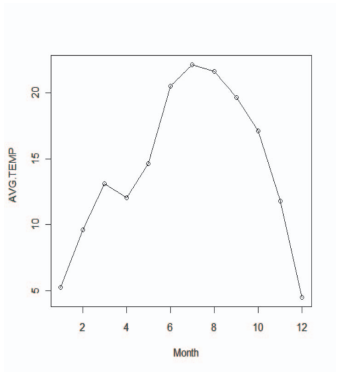
$$\mathrm{Temporal\ Variables} = \textit{Month of the year} + \textit{Day of the week} \tag{5}$$

$$\mathrm{Correlated\ measurement} = \textit{temperature} + \textit{humidity + wind + rain} \tag{6}$$

$$\mathrm{Classfires} = \{\text{accidental; small; medium; large}\}$$

D. Classifying the fire event hypothesis

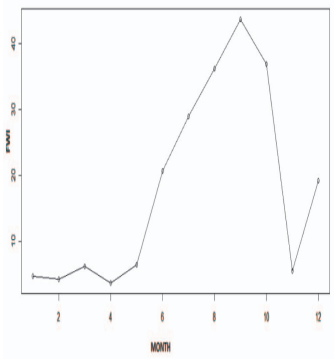
To train the original hypothesis with more available features we use the 517 rule sets from UCI forest fire



(a) Temperature temporal distribution.

FWI TYPES	Measure(norm)	Training set	Class μ
Low (AF)	0-8	0-8	31.7
Medium (SF)	8-13	8-13	33.2
High (MF)	13-32	13-39	33.4
Very High (LF)	$32 > FWI < 80$	$39 > FWI < 43$	36.8

(b) FWI class label thresholds.



(c) FWI temporal distribution.

Fig. 3. New classification of large fires based on $ISI = WIND + FFMC$ and $BUI = DMC + DC$ calculated as one FWI index.

repository. The weather attributes, which are numerical are converted into nominal [6,20] types, as given below.

$$temperature = \{cool; mild; hot\}$$

$$humidity = \{normal; high\}$$

$$wind = \{true; false\}$$

The model estimation of the target function with weights w_1, w_2 as shown allows to linearly separate in the feature space.

$$\hat{V} = w_1 x_1 + w_2 x_2 \tag{7}$$

The learning algorithm adjust the weights [6] for the misclassified samples using gradient decent techniques to find the local minima as shown in Figure 2(a) for sensor dataset. The the weights to minimizing the error and misclassifications as shown below.

$$E \equiv \sum_{i=0}^{m} (V_{train}(FireEvent) - \hat{V}(FireEvent))^2 \tag{8}$$

The expected new design of the classifier increases the classification accuracy of the baseline as shown in Table I from 62%to 84% in the case of small fire prediction. The test performance is marginal in terms of large fires with an accuracy of 7%.

V. Classification of a skewed distribution

In the previous section the small fires errors for normally distributed features could be reduced using higher dimensional feature space. In this section we try to learn events, which are temporal and also has a skewed distribution. The Canadian Fire Weather Index (FWI) [11] data allows to precisely predict large fires. FWI is calculated using Initial Spread Index(ISI) and Build Up Index(BUI), which take into account the fuel type such as pine accumulated in the ground. The fire index has two independent components, one which is the ground cover build up over time and the other the spread non-linear spread or fire risk factor as given in equation (9). Which indicate fire behavior and respectively represent rate of fire spread, fuel consumption and fire intensity and can be class labeled as given in Figure 3(b).

A. Temporal correlation of FWI features

All FWI indexes are significantly correlated with the number of fires and the burned area, specially when $BA > 100$ is the area classified as large fires. The average FWI index variation during the year and corresponding temperature is shown in Figure 3(a) and Figure 3(c), it increases during the month of May and peaks in August to September and starts reducing in the month of October. The target function defined in equation (3) can be defined for large fires as given in equation (9) in terms of Burnt area (BA). The first term is positively correlated as shown in Figure 2(b) and the second term spread index is non-linear in terms of fire activity.

$$BA_{FWI} > \textbf{50 ha} = (BUI) + (ISI)^x \tag{9}$$

Where Initial Spread Index (ISI), Buildup Index(BUI) are calculated from the weather logs [2] for a given fuel type and is provided in the UCI dataset. The FWI index is highly correlated with the number of fires and the burnt area. This can be seen in the plot in Figure 4(b) and 4(c), which shows $39 > FWI < 43$ (very high) in the case of large fires according to the training set during peak months. FWI class means are calculated and its distribution is multi-modal, which helps is accurate classification of large fires. The class means for large fires are shown in Figure 3(b) table.

B. F-score performance of FWI features

Large fires occurrence, which damage more than 50ha in total, amounts for majority of the burnt area(ha), it is a high priority to avoiding large fire incidents and help forest conservation. As they are hard to detect and

has a varying threshold it is also a cause of false alarms [16]. Plotting all the correlated FWI components, which relate to fire activity, Figure 4(c) shows that the peak months(Aug, Sep and Oct) has a gradual increase of FWI index and also correlated with large fire incidents. The area of high fire activity is shown in yellow, which has lowest false alarms rates for a given FWI threshold. The lower-bound conditions for fire activity is shown in blue, large fires in red and average FWI measured in green, which has high rates of false alarms due to unpredictable activity during the same time indicated in yellow.

The temporal attribute for large fire classification and its FWI F-score is plotted in Figure 2(c), it shows that large fires are invariant at F-score=0.9, and comparatively performs better than small fires with a relative score of (≤ 0.8). The expected testing accuracy of event classification will be close to optimal 95% performance.

VI. Generalization testing errors

Machine learning [7] algorithms, when used with density estimation and classification yields the lowest error. This allows to provide a baseline analysis of the measured attributes being used. Using Probably Approximately Correct (PAC)[9,19] learning for a hypothesis space $|H|$, which is finite then model can calculate the minimum number of training samples for a model with an accuracy $(1 - \delta_{train})$ and minimum allowable error ϵ_{train}. Its definition is given in Definition 2.

Definition 2 *Sensor networks trains on unique patterns, which are features present in the datastream, such attributes take real valued numbers. The performance of a system can be defined in terms of how well it can show the spatial and temporal changes of the data measured and classify as valid events. The learning algorithm depends on the distribution of the dataset and how many training sets are needed for a deterministic accuracy. Given the training sets is noiseless we use the PAC learning criteria given by, for some joint distribution $p(x, y)$,where x is the input variable and y represents the class label in, which class labels are determined by some function $y = g(x)$.*

$$E_{x,y}[I(f(x; \mathcal{D}) \neq y)] < \epsilon$$

A. Minimum number of samples for hypothesis testing

For a binary hypothesis of boolean literals the required training samples m is given by

$$|H|e^{\epsilon m} \leq \delta_{train} \tag{10}$$

For a decision tree the minimum number of training sets m of tree depth k is given by

$$m \geq \frac{69.0}{\epsilon_{train}} \left((2^k - 1)(1 + \log_2 H) + (1 + \log_2 \frac{1}{\delta_{train}}) \right) \tag{11}$$

From the given dataset, we can calculate m from equation (11) by substituting $k = 4$ as we have four classes,$|H| = 9$, the average accuracy $(1 - \delta_{train} = 0.8)$ and $\epsilon_{train} = 0.2$. A 20% error rate is chosen due to unknown distribution.

$$m \geq \frac{69.0}{0.2} \left((2^4 - 1)(1 + \log_2 9) + (1 + \log_2 \frac{1}{0.2}) \right)$$

Then the minimum number of training support vectors [18] can be computed as given below.

$$m_{SV} \geq 212$$

Definition 3 *Measuring a sensor value $\bar{x}_i$ exactly or within a small error ϵ_{sample} and penalty of 1 for not getting the correct value. This leads to the $0 - 1$ loss function, which is defined by*

$$L|\theta, x| = \begin{cases} 0 & \text{if } |x - \theta| \leq \epsilon_{sample} \\ 1 & \texttt{otherwise} \end{cases}$$

We choose θ to minimize the function, for a given dataset it can be shown that it is the mode.

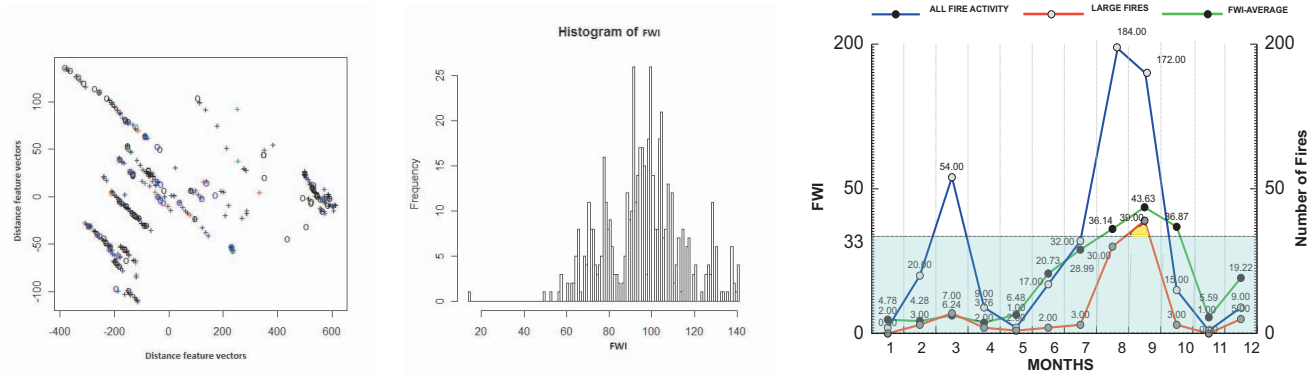
$$R\theta = E|L(\theta, x] = \sum_{i=0}^{n} P(x_i)$$

$$= 1 - \sum_{i=0}^{n} P(x_i)$$

Definition 4 *In practice errors are high in the case of sensor data, due to the class-conditional distributions overlaps, in which case exact separation of the training data can lead to poor generalization. We therefore need a way to modify the support vector machine so as to allow some of the training points to be misclassified. To do this, we introduce slack variables, $\xi_n \geq 0$ where $n = 1, ..., N$, with one slack variable for each training data point [20]. These are defined by $\xi = 0$ for data points that are on or inside the correct margin boundary and $\xi = |tn - y(x_n)|$ for other points. Thus a data point that is on the decision boundary $y(xn) = 0$ will have $\xi = 1$, and points with $\xi > 1$ will be misclassified. The exact classification constraints [19] are then replaced with $t_n y(x_n) \geq 1 - \xi_n, \quad n = 1, ..., N$, in which the slack variables are constrained to satisfy $\xi_n = 0$. Data points for which $\xi_n = 0$ are correctly classified and are either on the margin or on the correct side of the margin. Points for which $0 < \xi_n < 1$ lie inside the margin, but on the correct side of the decision boundary, and those data points for which $\xi_n > 1$ lie on the wrong side of the decision boundary and are misclassified. The goal is now to maximize the margin, while softly penalizing points that lie on the wrong side of the margin boundary. We therefore minimize the equation*

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2} ||w||^2$$

where C is the loss-function of the SVM.



(a) Support vectors shown for SVM classifier. (b) FWI Multi-modal Density plot. (c) FWI<= 39(blue) and FWI>39(yellow).

Fig. 4. Positively correlated feature performance, shows all large fires(red) fall into the yellow region.

Other Test	Yellow Region	Blue Region	Temporal-Small fires	FWI-Large fires
Alarm	H=Hit	H=Hit	H=311	H=69
No Alarm	M=Miss	M=Miss	M=111	M=3
False Alarm	F=False Alarm	Z=Null hypothesis	F=8	F=307

Table Alarms: Performance of classification with false alarm rates.

B. Misclassification cost of false alarms

While testing the performance of the machine learning [7] algorithm, confusion matrix gives the error rates of misclassified samples, which allows to find the sensitiveness of the system to false positives. False positives have more significance when detecting large fires, which are very rare and hard to detect. A cost function [8] can be used, which is shown in the matrix below. The cost function penalizes misclassification of large fires. We define a hit count in terms of precision to avoid false alarms [16]. From the alarms Table part of Figure 4, the results shows that when using FWI calculation the alarms are very reliable for large fires and any misclassification is penalized. While the false alarms are higher for small fire detection the penalty of misclassification is lower by design assumptions.

$$\begin{pmatrix} & Decision\ \ made & \\ & LargeFire & SmallFire \\ LargeFire & 0 & 1000 \\ SmallFire & 1 & 0 \end{pmatrix}$$

Cost function matrix for misclassification

VII. SIMULATION

Open-source workbench called WEKA [3] is a useful tool to quantify and validate results, which can be reliably duplicated. WEKA can handle numeric attributes well, so we use the same values for the weather data from the UCI [4] repository datasets. The popular classification algorithms are cross-validated with a subset of the data from the original dataset and the results are discussed in this section. The error minimization for the error generalization to avoid over-fitting the design is discussed in Definition 1,2,3 and 4. The results can be extended with real test data sets acquired from sensors samples.

A. Data Mining analysis

In the pre-processing of the dataset, we use a high F-score to design the classifier for better performance and the training and testing errors are designed according to PAC learning and sensor sampling as described in section VII. The results from Table I shows training and testing set accuracy using sensor data, small fires have more accuracy with a higher bound performance of $(1-\delta_{train} = 95\%)$ and lower bound of $(1-\delta train = 30\%)$. This results is consistent with the F-score ranking analysis. The true error is calculated for all the classes using cross-validation of the training set, the performance is very marginal $1-\delta_{test} = 56\%, \epsilon_{sample-noise} = 20\%$ as the distribution of all classes are not normally distributed and prone to statistical errors [10], the attribute distribution is shown in Figure 3 and Figure 4.

B. FWI analysis

The second dataset uses the new standards established by the Forest fire department as shown in Figure 3(b) with redefined class labels. As seen earlier the majority of training errors are large fire detection, where the sensor data is not accurately estimated. The FWI index performs well due to the features being positively correlated. The results from Table II shows that training accuracy has increased for hard to classify fires from the previous case. We get $(1-\delta_{train} = 95\%)$ classification accuracy with 417 support vectors, when using a polynomial kernel. This result is also consistent with the F-score ranking of FWI features.

C. Cost function analysis

In an effort to deal with outliers in the form of false alarms a cost function is used when testing. The cost function definition allows to lower 310 false alarms to only 3 for the FWI dataset, it still has testing error in

	FWI Model Parameters				Alarm Accuracy					
Experiments	**Features**				**Training Error**					**True Error $\epsilon_{outlier}$**
	Class	Attr.	Performance	SV	AF	SF	MF	LF	err	Outliers
Hypothesis	4	3	F-measure	517	45%	45%	90%	90%	4%	-
Hypothesis	2	3	Subset	95	-	-	95%	95%	5%	-
SVM	1	1	FWI	517	-	-	97%	100%	4%	72%

Class=Multi-class Attr=No. of attributes SV=Support Vectors AF=Accidental Fire SF=Small Fire MF=Medium Fire LF=Large Fire err=Misclassification Outlier=False alarms

TABLE II
F-MEASURE PERFORMANCE FOR ALL TESTS USING FWI ATTRIBUTES.

terms of outliers, which accounts for $\epsilon_{outlier} = 72\%$. The reliability of the false alarms as discussed in Definition 4 rates is consistent with FWI features ranking as it is invariant to any other features part of the dataset.

VIII. Summary

In this work we discuss a knowledge framework for tracking weather phenomenon and rare fire events. The framework uses machine learning ranking using F-score, data mining logs from automated weather sensors and cost function to allow non-linearity in the form of false alarms for determining the accuracy and allowable errors of a event detection classifier. We effectively solve the problem of unattended classification and outlier detection by using sensor data and define a training framework for the user to query the measured values within an acceptable error. The performance of the classifier shows that the classification accuracy can be increased by adding inexpensive sensors to the original dataset and small fires can be reliably predicted. Using FWI dataset helps the fire fighting personal during the peak fire months by accurately predicting the possibility of large forest fires. Using inexpensive sensors in remote areas can further help in knowledge discovery and learning event patterns and classification. A user defined cost function allows to choose the sensitiveness of the false alarms and practically help the testing of the new classifier before deployment.

IX. acknowledgement

One of the authors like to thank Shailesh Kumar of Google, Labs, India for suggesting the machine learning framework WEKA and related topics in statistical methods in Data mining. Authors like thanks all the encouragement from IIIT research community in Hyderabad on this data mining project and specially Dean P.J. Narayanan, Dr. Anoop, Dr. Jawahar and Dr. Parekh. The first author likes to express appreciation and support from the advisors Dr. S.S Iyengar, Dr. Rama Murthy and Dr. Rawat in this research effort, which was funded under the LSU, Baton Rouge research grant. Authors also are grateful and like to thank the anonymous reviewer's comments, which has improved the final quality of the submission.

References

[1] Vasanth Iyer, S.S. Iyengar, N. Paramesh, G. Rama Murthy, and M.B. Srinivas. Machine Learning and Data Mining Algorithms for Predicting Accidental Small Forest Fires. SENSORCOMM 2011, August 21-27, 2011 - French Riviera, France.

[2] Paulo Cortez and Anibal Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. Department of Information Systems-R&D Algoritmi Centre, University of Minho, Portugal.

[3] WEKA Machine learning software. $http : //www.cs.waikato.ac.nz/\ ml/weka$ [Accessed May 15th, 2011].

[4] Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. [Accessed May 20th, 2011].

[5] Vasanth Iyer, S. Sitharama Iyengar, Garmiela Rama Murthy, N. Parameswaran, Dhananjay Singh, and Mandalika B. Srinivas. Effects of channel SNR in Mobile Cognitive Radios and Coexisting Deployment of Cognitive Wireless Sensor Networks. IEEE IPCCC, 2010, pp. 294-300. Albuquerque, New Mexico, USA.

[6] Ian H. Witten and Eibe Frank. Data Mining, Pratical machine learning. Elsevier 2005.

[7] Tom M. Mitchell. Machine Learning. MaGRAW-Hill Publications 1997.

[8] Peter L. Bartlett and Marten H. Wegkamp. Classification with a Reject Option using a Hinge Loss. Journal of

[9] David Hawking, and Stephen Robertson. On Collection Size and Retrieval Effectiveness. CSIRO mathematical and Information Sciences, Canberra, Australia.

[10] Mark D. Smuker, James Allan, and Ben Carterette. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. Department of Computer Scince, University of Massachusetts, Amherst.

[11] Using the Canadian Fire Weather Index (FWI) in the Natural Park of Montesinho, NE Portugal: calibration and application to fire management.

[12] Shailesh Kumar, Joydeep Ghosh, and Melba M. Crawford, IEEE member. Best-Bases Feature Extraction Algorithms for Classification of Hyperspectral Data.

[13] Raymond Wan, and Alistair Moffat. Interactive phrase browsing within compressed text. omputer Science and Software Engineering, University of Melbourne. 2001.

[14] H. Harb, and L. Chen. A Query by example music retrieval algorithm. Maths-Info department, Ecole Centrale de Lyon. France, 2001.

[15] Richard R. Brook, and S. S. Sitharama Iyengar. Robust Distributed Computing and Sensing Algorithm, ACM, 1996.

[16] Bhaskar Krishnamachari, and S.S Iyengar. Distributed Bayesian Algorithms for Fault-Tolerant Event Region Detection in Wireless Sensor Networks, IEEE Transaction on Computers, Vol. 53, No. 3, 2004.

[17] C. Cortes and V. Vapnik. Support-vector network. Machine Learning, 20:273-297, 1995.

[18] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer 2006.

[19] C.J. van Rijsbergen. Information Retrieval. Butterworths, London, second edition, 1979.

[20] Zadeh, L., Fuzzy sets, Information Control 8, 338-353, 1965