International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

EFFICIENT FEATURE SELECTION FOR FAULT DIAGNOSIS OF AEROSPACE SYSTEM USING SYNTAX AND SEMANTIC ALGORITHM

Meena E[1], Revathi B[2],
Sajanvethakumar F[3](Assistant professor of Computerscience and Engineering)

[123]*Department of Computer Science and Engineering, JEPPIAAR SRR Engineering College, Padur, Chennai 603103*

ABSTRACT:*Each and every year, the Aerospace system handles the fault verbatim record database. So the usage of fault verbatim record database is to generate the fault by text, if the airplane does not pass the signal code at correct time when the Airplane starts. It has high dimensional data, learning difficulties and with unstructured verbatim record. Learning difficulties, if the person have little amount of English knowledge, it find difficult to understand. High dimensional data, if the fault having 3 to 4 lines then it may take some*

time to understand and identify the faults. In proposed system we introduce, Bi-level Feature Extraction Based Text Mining. Bi-level is nothing but the comparison of higher order and lower order. It fault feature derived from both syntax level and semantic level. Syntax level used to overcome the learning difficulties and the semantic level use to convert high dimensional to the low dimensional. It can be used to diagnosis the problem quickly and rectify the problems.

1. INTRODUCTION

Text mining could be a knowledge-intensive task and is gaining a lot of and a lot of attention in many industrial fields, as an example, aerospace, automotive, railway, power, medical, biomedicine, producing, sales and selling sectors. In a railway field, advanced data technologies, such as sensing element networks, RDIF techniques, wireless communication, and net cloud, area unit won't to monitor the health of the aerospace systems. In the event of malfunctioning, the diagnostic hassle symptoms are generated and transmitted to the watching center info by wired/wireless communications. When each diagnosis episode a repair verbatim is recorded, that consists of a matter description of the mixture of fault symptom (i.e., fault terms), e.g., "Speed Distance Unit (SDU) relevant faults," a fault symptom e.g., "SDU," failure modes (i.e., fault classes), and at last corrective actions, e.g., "replaced SDU," taken to repair its faults.

However, the task of automatic discovery of information from the repair verbatim may be a non-trivial exercise primarily owing to the following reasons:

1) High-dimension information. In maintenance documents, there are tens of thousands or maybe many thousands of distinct terms or tokens. when elimination of stop words and stemming, the set of options continues to be overlarge for many learning algorithms.

2) unbalanced fault category distribution. In maintenance documents, the number of examples in one fault category (i.e., majority class) is considerably larger than that of the others (i.e., minority classes). Such unbalanced category distributions have exhibit a heavy issue to most classifier learning algorithms that assume a comparatively balanced distribution.

3) unsupervised text mining models. They will not turn out topics that adjust to the user's existing information. One key reason is that the target functions of topic models, e.g., Latent Dirichlet Allocation, LDA , typically don't correlate well with human judgments.

This work proposes a bi-level feature extraction-based text mining for fault designation to fulfill the aforesaid challenges by mechanically analyzing the repair verbatim. Our main plan is to extract fault options at syntax and linguistics levels severally so fuse them to realize the required results. Considering the very fact that the extracted options at every level offers a distinct stress to a specific facet of feature spaces and has its deficiencies, the planned feature fusion of two levels could enhance the exactness of fault designation for all fault categories, particularly minority ones.

At the syntax level, we have a tendency to propose associate degree improved $\chi 2$ statistics (ICHI) to deal with the feature choice of unbalanced information set. First, we have a tendency to overcome the negative result of unbalanced information set by adjusting the feature weight of minority and majority classes. This makes minority categories comparatively distant from the majority ones. Second, we have a tendency to contemplate the Hellinger distance as a choice criterion for feature

choice, which is shown to be imbalance-insensitive. The planned ICHI may be regarded as feature picks at the syntax level as a result of it mainly uses the document-word matrix. At the linguistics level, we have a tendency to borrow the thought from and propose an LDA with previous data (ab. PLDA) to perform the feature extraction. By representing documents in topics rather than word house, we have a tendency to area unit able to offer additional feature extraction at the linguistics level to compensate those extracted at the syntax level. the mixing of previous data with the fundamental LDA is based on the very fact that LDA, as associate degree unattended model, cannot deal with such problems as choosing topic counts and reducing the adverse result of common words, which cannot turn out topics that adapt to a user's existing data. Previous data helps U.S.A. guide topic mining in basic LDA. Finally, we have a tendency to fuse the extracted options derived from the syntax level with the linguistics one by serial fusion to boost Support Vector Machine (SVM)-based fault diagnosing for all fault categories, particularly minority ones.

2. RELATED WORK

To manage the challenges obligatory by unbalanced category distributions, several learning algorithms are planned. For instance, the sampling-based strategies, e.g., over-sampling scheme and under-sampling theme square measure the best yet effective ones, within which categories square measure replicated or curtail to achieve an identical balanced result. Another well-liked methodology is the value-sensitive learning theme that takes the price matrix into thought throughout model building and generates a model that has all-time low value. Margineantu et al. examined various strategies for incorporating value data into the C4.5 learning formula. Joshi et al. planned PNrule, a two-phase rule induction formula, to handle the mining of minority classes. Tang et al. incorporated completely different rebalance heuristics, as well as cost-sensitive learning, over-sampling and under-sampling in SVM modeling and introduced four SVM variations to tackle the imbalance learning downside. A survey about this subject is found in Mladenic et al. discussed the feature choice problems for unbalanced category distributions. However, this work is restricted to the Naive Bayesian classifier. Also, Zheng et al. planned a feature choice method for unbalanced text documents by adjusting the mix of positive and negative options within the information. Their method sticks to the normal goodness measures of options. Yin et al. planned to divide the bulk category into comparatively smaller pseudo-subclasses with comparatively uniform sizes to manage influence of unbalanced information sets. In text mining-based feature extraction, applied math and graphic modeling has been paid a lot of and a lot of attention and thought of as a well-liked and economical tool to mine topics to scale back dimensions. For example, LDA was antecedently wont to construct features for classification. It usually acts to scale back information dimension. In distinction, the essential LDA, as AN unattended model, cannot perform to an adequate degree during a topic mining method. To solve this downside, Andrzejewski et al. incorporated domain information by employing a Dirichlet Forest previous in LDA. Zhai et al. planned probabilistic constraints as a relaxation mechanical modification, that could be a soft constraint, to the chemist sampling equation. Hospedales proposed weakly supervised joint topic model that learned a model for all the classes by employing a part shared common basis. Wang proposed a unnatural topic model by adding constraints to guide a subject mining method, that improved the accuracy of mining topics.

3. ICHI-BASED FEATURE SELECTION AT SYNTAX LEVEL

The basic idea of the proposed ICHI is to make a minority class far away from the majority one by adjusting weights of fault terms as shown in Fig. 1. To facilitate understanding, we first define some notations. Tm is the set of fault terms of minority fault classes, TM the set of fault terms of majority fault class and Tc, the intersection of Tm and TM, the common feature set.

SYNTAX LEVEL ALGORITHM

Data: Dataset S, fault term T, fault class F
Result: Feature set F1
Begin
W← word segmentation
M← word-Document matrix
For w_i Є W and f_j Є F do
R(i,j)← correlation between fault term and class
End
R1← normalization of R
F1(i)←Fault feature
For f_i, f_j Є F do
F2(i,j)← common fault feature set of fault class by intersection of feature set
End
F2← common fault feature set by union
For f_i Є F do

F2(i)← Exclusive feature set by excluding F2

W1(i)← Weight of F2(i) by inverse probability

End

For w_k Є F do

L(w_k)← Hellinger distance

F1(i,j)← common feature set selected by highest k features according to hellinger distance L

End

End

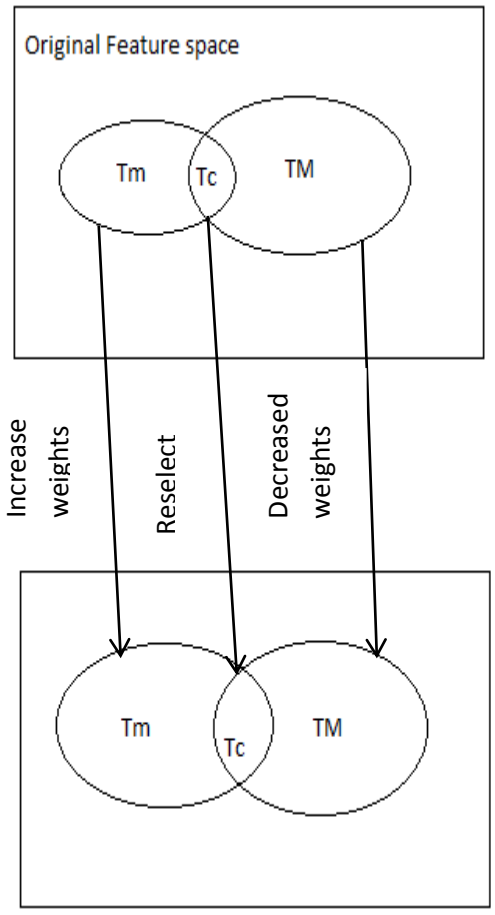


Fig1. Idea of proposed ICHI

Let image / denote the set distinction, Tm/Tc and TM/Tc square measure related with minority and majority categories solely, severally, thereby known as them as exclusive fault term sets.

3.1 χ2 Statistics and Hellinger Distance

χ2 statistics could be used to estimate the shortage of independence between a term t and a class ci and might be compared to the χ2 distribution with one degree of freedom to evaluate extremeness. It's outlined as:

$$\chi^2(t,c_i) = \frac{N[P(t,c_i)P(\bar{t},\bar{c}_i)-P(t,\bar{c}_i)P(\bar{t},c_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)} \quad (1)$$

where N is that the total number of documents. (t, c_i) denotes the presence of term t and its membership in class c_i, $P(t,\bar{c}_i)$presence of t however not its membership in ci, (t, ci) absence of t but its membership in ci, and $P(\bar{t},\bar{c}_i)$ absence of t and its nonmembership in c_i. $P(\cdot,\cdot)$ means that the likelihood of presence/absence of term t and its membership/non-membership in class c_i.

Hellinger distance may be a live of spatial arrangement divergence. Given 2 separate likelihood distributions P = {p1,p2,..pn} and Q={q1,q2,...qn}, their Hellinger distance is outlined as:

$$H(P,Q) = \frac{\sqrt{1}}{2}\sqrt{\sum_i^k(\sqrt{pi} - \sqrt{qi})2} \quad (2)$$

By definition, the Hellinger distance may be a metric satisfying triangle difference.√2 within the definition is employed for making certain that H(P,Q) ≤ one for all likelihood distributions.

3.2 ICHI Based Feature selection at Syntax Level

The main steps of ICHI-based feature choice area unit summarized by algorithmic program one. once a fault maintenance document D and a fault term wordbook Ω area unit provided, word set W (i.e., fault term set) is extracted by word segmentation.

According to W and fault categories C, a word-document matrix M can be generated (lines 1-2). Then we have a tendency to cypher correlations R between feature terms and fault categories by χ2 statistics (lines 3-4). so as to check the correlation between totally different fault terms and totally different categories, we have a tendency to normalize them as follows (line 5):

$$R(w_i, c_j) = \frac{R(i,j)}{\sum_{i=1:m} R(i,j)} \times \frac{R(i,j)}{\sum_{j=1:n} R(i,j)} = \frac{R(i,j)^2}{\sum_{i=1:m} R(i,j) \times \sum_{j=1:n} R(i,j)} \quad (3)$$

where n is that the variety of fault terms contained in W, m is the number of fault categories in C. In Eq. (3), the correlation of feature term Badger State and fault category c_j depends on the correlations between term Badger State and every one different fault categories besides c_j . Therefore, it is depicted exactly by the merchandise of R(i, j)/∑i=1:mR(i, j) and R(i, j)/∑j=1:nR(i, j). we have a tendency to then choose highly connected fault feature sets F for every fault category by comparing correlations with a given threshold (line 6). Next, lines 7–9 acquire the inclined fault feature set $\hat{F}$ by intersecting each combine of fault term sets. At an equivalent time, the exclusive feature sets $\tilde{F}$ of every fault category is obtained in line twelve. Next, we have a tendency to change their weights in step with chances of their corresponding fault categories (line 13).

To the gravity fault term set $\hat{F}$, we want to judge the distributive discrimination of every feature on fault categories by computing its Hellinger distance with these fault categories victimization Eq. (2) (line 16). Then we have a tendency to use it to reselect the common options of each fault category pairwise (line 17). At last, we have a tendency to get the ultimate common feature set ($\hat{F}'$) of the information set by performing arts the union of all the common feature sets of all fault categories pairwise (line 19). Thus, we have a tendency to complete the feature choice of fault term features and find such feature space F_a as [(exclusive feature sets, weights), common feature set] (line 20), i.e., [($\tilde{F}$, $\tilde{F}$),$\hat{F}_{\varpi}$].

4. PLDA BASED FEATURE SELECTION AT SEMANTIC LEVEL

In this section, we first get to know about LDA and so introduce the extraction of relationship supported previous information. At last we have a tendency to gift the projected PLDA that comes with prior information into LDA to appreciate the feature choice at the semantic level.

SEMANTIC ALGORITHM

Data: Dataset S, Fault class F, Topic sets K
Result: Correlation $r(w_i,z_k)$

Begin
R1← Normalization of R
Ξ← k clusters
Θ← degree of correlation
For w_i Є W and f_i Є F do
If $R1(w_i,f_j)$ is highest or lowest two ranks in Ξ then
$R1(w_i,f_j)$is assigned SR or WR
Else
$R1(w_i,f_j)$is assigned as CR
End
End
Fault classes f_i Є F is preassigned with two corresponding copies z_{2*I}, z_{2*i+1}
r (w_i,z_k)←initialize correlation between term and topic with zeros
For w_i Є W and z_k Є Z do
If z_k Є f_j then
(w_i,z_k) is assigned with the value of $R1(w_i,f_j)$
End
End
End

4.1 LDA

Given D documents expressed over W distinctive words and T topics, LDA outputs the document-topic distribution and topic word distribution, each of which may be obtained with chemist Sampling. Its key step is that the topic change for every word in every document in step with

$$P(z_i=j|z_{-i}, w, \alpha, \beta) \propto \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(di)} + \alpha}{n_{-i,\cdot}^{(di)} + T\alpha} \quad (4)$$

where $z_i=j$ denotes the ith word in an exceedingly document appointed to topic j, z_{-i} all the subject assignments apart from the ith word, i.e., the current one. w= {w_1,w_2,w_3,....w_n}, wherever every W_i belongs to some document.α and β are hyper-parameters for the document-topic and topic-word Dirichlet distributions, severally $n_{-i,j}^{(wi)}$ is that the total range of same words Wisconsin

appointed to topic j, not together with this one and $n^{(.)}_{-i,j}$ the full range of words appointed to topic j, not together with this one. $n^{(di)}_{-i,j}$) is the range of words from document di appointed to topic j, not together with this one, and $n^{(di)}_{-i,.}$ is that the total range of words in document di, excluding this one. After M iterations of chemist sampling for all words altogether documents, the distribution φ and θ are finally calculable as follows:

$$\varphi_j^{(wi)} = \frac{n_j^{wi} + \beta}{\sum_{w'}^{W} n_j^{wi} + W\beta} \quad (5)$$

$$\theta_j^{(di)} = \frac{n_j^{di} + \alpha}{\sum_{j}^{T} n_j^{di} + T\alpha} \quad (6)$$

4.2 Extraction of Relationship-Based data

To facilitate understanding the extraction of previous data, we offer 3 varieties of relationship between fault terms and fault classes.

Strong Relationship (SR): fault terms powerfully relate with a specific fault category and hardly relate with others. Hence PLDA adds these options to the precise fault category in topic mining based fault choice.
Weak Relationship (WR): fault terms hardly relate with a specific fault category. These fault terms shouldn't be associated with the precise fault category.
Complex Relationship (CR): fault terms powerfully relate with more than one fault category. we must always provide it comprehensive considerations in topic mining-based fault choice.

The main steps of previous data extraction are summarized into semantic algorithm. Like syntax algorithm the normalized correlations ($\bar{R}$) is calculated by Line one. Then $\bar{R}$ is clustered into eight clusters Ξ by the K-means bunch methodology (Line 2). Correlation degree (Θ) between fault terms and fault categories, such as SR, WR and CR, is then assigned to every pairwise term and fault category (Lines 4–8). During this work, every fault category is pre assigned with 2 corresponding topics. as an example, topics z_{2*i}, z_{2*i+1} corresponds fault $c_i \in C$ ($1 \le i \le |C|$), where |C| represents category count. Then the correlation (Γ) between terms and topics will be obtained (lines 13–15).

4.3 Incorporating previous data Into LDA

The main plan of incorporating previous data into LDA is to revise the subject change possibilities by victimization previous information. That means, during a topic change method in (4), we multiply an extra indicator operate $\delta(w_i, z_j)$, which represents a tough constraint of SR and WR from terms to topics.
The final probability for topic change is:

$$P(z_i = j|z_{-i}, w, \alpha, \beta) \propto \delta(w_i, z_j) * \frac{n^{(wi)}_{-i,j} + \beta}{\sum_{w'}^{W} n^{wi}_{-i,j} + W\beta} \frac{n^{(di)}_{-i,j} + \alpha}{\sum_{j}^{T} n^{di}_{-i,j} + T\alpha} \quad (7)$$

where $\delta(w_i, z_j)$ represents intervention or facilitate from pre-existent knowledge of SR and WR, that plays a key role in this update. Within the topic change {for every|for every} word in each document, $\delta(w_i, z_j)$ equals $\Gamma(w_i, z_j)$. For advanced relationship (CR), influence of fault term Badger State and fault categories on topic-word distribution ought to be all taken into account. Our basic plan is to see the association between w_i and C_{zj}, wherever C_{zj} denotes the set of fault categories to that topic z_j hooked up. If they have relevance higher than a pregiven threshold, $\Gamma(w_i, z_j)$ ought to be assigned a positive variety. Otherwise, $\Gamma(w_i, z_j)$ is set as a negative variety. Therefore, (4) is revised as follows:

$$P(z_i = j|z_{-i}, w, \alpha, \beta) \propto \frac{(1 + F_{wi,zj})\, n^{(wi)}_{-i,j} + \beta}{\sum_{w}^{W}(1+F_{w,zj})\, n^{w'}_{-i,j} + W\beta} \frac{n^{(di)}_{-i,j} + \alpha}{n^{(di)}_{-i,.} + T\alpha} \quad (8)$$

where $F_{wi,j}$ corresponds to $\Gamma(w_i, z_j)$in semantic algorithm and reflects the correlation of fault term w_i with topic z_j. Then (8) is used to modification the sampling method for fault knowledge set with CR relationship.

5. SERIAL FAULT FEATURE FUSION

The fault feature extracted at the syntax level is united with those at the linguistics level. To facilitate understanding, we denote the processed fault feature from the syntax level as $F_a = (a_1, a_2, \ldots, a_M)$ and also the one from

linguistics level $F_b = (b_1, b_2, \ldots, b_N)$, wherever M and N square measure the dimension at syntax and linguistics levels severally. Here we tend to adopt a serial fusion method to make a combined feature F_γ. it's outlined by

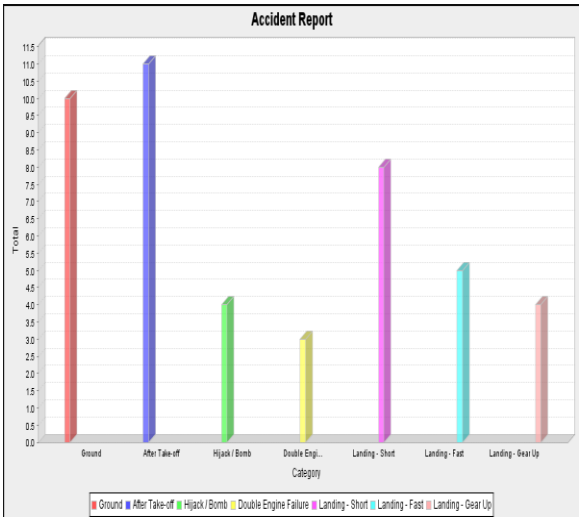
$$F_\gamma = (F_a, \theta * F_b) = (a_1, a_2, \ldots, a_M, \theta * b_1, \theta * b_2, \ldots, \theta * b_N) \quad (9)$$

where θ is associate adjusting parameter. It may be obtained from training set through learning. once the accuracy modification in 2 continuous iterations is a smaller amount than 0.1, we tend to set this price as θ. All serially combined feature vectors kind associate (M+N)-dimensional feature space.

6. EXPERIMENTAL RESULTS

The main cause of the accidents shows the following results

1. Ground
2. After Take-off
3. Hijack / Bomb
4. Double Engine Failure
5. Landing - Short
6. Landing - Fast
7. Landing - Gear Up



7. CONCLUSION

Text mining of repair verbatim for fault diagnosis of Aerospace systems poses a big challenge due to unstructured verbatim, high-dimension data, and imbalanced fault classes. In this paper, to improve the fault diagnosis performance, especially on minority fault classes, we have proposed a bi-level feature extraction-based text mining method. We first adjust the exclusive feature weights of various fault classes based on χ2 statistics and their distributions. Then we reselect the common features according to both relevance and Hellinger distance. This can be categorized as feature selection at the syntax level. Next, we extract semantic features by using a prior LDA model to make up for the limitation of fault terms derived from the syntax level. Finally, we fuse fault term sets derived from the syntax level with those from the semantic level by serial fusion.

REFERENCES

[1] L. Huang and Y. L. Murphey, "Text mining with application to engineering diagnostics," in *Proc. 19th Int. Conf. IEA/AIE*, Annecy, France, 2006, pp. 1309–1317.
[2] D. G. Rajpathak, "An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain," *Comput Ind.*, vol. 64, no. 5, pp. 565–580, Jun. 2013.
[3] J. Silmon and C. Roberts, "Improving switch reliability with innovative condition monitoring techniques," *Proc. IMechE, F C J. Rail Rapid Transit*, vol. 224, no. 4, pp. 293–302, 2010.
[4] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[5] J. Chang, J. Boyd-Graber, C.Wang, S. Gerrish, and D. Blei, "Reading tea leaves: How humans interpret topic models," *Neural Inf. Process. Syst.*, vol. 22, pp. 288–296, 2009.
[6] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanced data," in *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases-Part I*. Berlin, Germany: Springer-Verlag, 2008, pp. 241–256.
[7] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. 15, no. 1, pp. 52–60, Feb. 1967.