IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems

D.Keerthanaa[1*], C. Premila Rosy[2]

[1]Department of Computer Science, Idhaya College for Women, Kumbakonam, Tamilnadu, India
[2]M.Sc Computer Science, Idhaya College for Women, Kumbakonam, Tamilnadu, India

Corresponding Author: premilarosy78@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—**A vast amount of text data is recorded in the forms of repair verbatim in railway maintenance sectors. Efficient text mining of such maintenance data plays an important role in detecting anomalies and improving fault diagnosis efficiency. However, unstructured verbatim, high-dimensional data, and imbalanced fault class distribution pose challenges for feature selections and fault diagnosis. We propose a bilevel feature extraction-based text mining that integrates features extracted at both syntax and semantic levels with the aim to improve the fault classification performance. We first perform an improved $\chi 2$ statistics-based feature selection at the syntax level to overcome the learning difficulty caused by an imbalanced data set. Then, we perform a prior latent Dirichlet allocation-based feature selection at the semantic level to reduce the data set into a low-dimensional topic space. Finally, we fuse fault features derived from both syntax and semantic levels via serial fusion. The proposed method uses fault features at different levels and enhances the precision of fault diagnosis for all fault classes, particularly minority ones. Its performance has been validated by using a railway maintenance data set collected from 2008 to 2014 by a railway corporation. It out performs traditional approaches.

***Keywords*—**Bilevel, Feature Selection, Feature Extraction, Railway, Text Mining.

I. INTRODUCTION

a) INTELLIGENT TRANSPORTATION SYSTEMS

Intelligent transportation systems (ITS) are advanced applications which, without embodying intelligence as such, aim to provide innovative services relating to different modes of transport and traffic management and enable various users to be better informed and make safer, more coordinated, and 'smarter' use of transport networks[1] – [5]. Although ITS may refer to all modes of transport, defined ITS as systems in which information and communication technologies are applied in the field of road transport, including infrastructure, vehicles and users, and in traffic management and mobility management, as well as for interfaces with other modes of transport.

b) INTELLIGENT TRANSPORTATION TECHNOLOGIES

Intelligent transport systems vary in technologies applied, from basic management systems such as car navigation; traffic signal control systems; container management systems; variable message signs; automatic number plate recognition or speed cameras to monitor applications, such as security CCTV systems; and to more advanced applications that integrate live data and feedback from a number of other [6] [8] [9]sources, such as parking guidance and information systems; weather information; bridge de-icing (US deicing) systems; and the like. Additionally, predictive techniques are being developed to allow advanced modelling [7] [10] and comparison with historical baseline data. Some of these technologies are described in the following section.

c) WIRELESS COMMUNICATIONS

Various forms of wireless communications technologies have been proposed for intelligent transportation systems. Radio modem communication [11] on UHF and VHF frequencies are widely used for short and long range communication within ITS.Short-range communications [13] [15] [12] of 350 m can be accomplished using IEEE 802.11 protocols, specifically WAVE or the Dedicated Short Range Communications standard being promoted by the Intelligent Transportation Society of America and the United States Department of Transportation. Theoretically, the range of these protocols can be extended using Mobile ad hoc networks or Mesh networking. Longer range communications have been proposed using infrastructure networks such as WiMAX (IEEE 802.16), Global System for Mobile Communications (GSM), or 3G. Long-range

communications using these methods are well established, but, unlike the short-range protocols [14] [16] [17] 18], these methods require extensive and very expensive infrastructure deployment. There is lack of consensus as to what business model should support this infrastructure.

d) COMPUTATIONAL TECHNOLOGIES

Recent advances in vehicle electronics have led to a move towards fewer, more capable computer processors on a vehicle. A typical vehicle in the early 2000s would have between 20 and 100 individual networked microcontroller/Programmable logic controller modules with non-real-time operating systems. The current trend is toward fewer, more costly microprocessor modules with hardware memory management and real-time operating systems. The new embedded system platforms allow for more sophisticated software applications to be implemented, including model-based process control, artificial intelligence, and ubiquitous computing. Perhaps the most important of these for Intelligent Transportation Systems [19] [20] is artificial intelligence.

e) FLOATING CAR DATA/FLOATING CELLULAR DATA

In developed countries a high proportion of cars contain one or more mobile phones. The phones periodically transmit their presence information to the mobile phone network, even when no voice connection is established. In the mid-2000s, attempts were made to use mobile phones as anonymous traffic probes. As a car moves, so does the signal of any mobile phones that are inside the vehicle.

II. METHODOLOGY

At the syntax level, we propose an improved $\chi 2$ statistics (ICHI) to cope with the feature selection of imbalanced data set. First, we overcome the negative effect of imbalanced data set by adjusting the feature weight of minority and majority classes. This makes minority classes relatively far away from the majority ones. Second, we consider the Hellinger distance as a decision criterion for feature selection, which is shown to be imbalance-insensitive. The proposed ICHI can be regarded as feature selections at the syntax level because it mainly uses the document-word matrix.

input : positive documents D^+; minimum support, *min_sup*.
output: d-patterns DP, and supports of terms.

$DP = \emptyset$;
foreach *document* $d \in D^+$ **do**
  let $PS(d)$ be the set of paragraphs in d;
  SP = SPMining($PS(d)$, *min_sup*);
  $\hat{d} = \emptyset$;
  foreach *pattern* $p_i \in SP$ **do**
    $p = \{(t, 1) | t \in p_i\}$;
    $\hat{d} = \hat{d} \oplus p$;
  end
  $DP = DP \cup \{\hat{d}\}$;
end
$T = \{t | (t, f) \in p, p \in DP\}$;
foreach *term* $t \in T$ **do**
  $support(t) = 0$;
end
foreach *d-pattern* $p \in DP$ **do**
  foreach $(t, w) \in \beta(p)$ **do**
    $support(t) = support(t) + w$;
  end
end

III. RESULTS AND DISCUSSION

This section presents the results for the evaluation of the proposed approach PTM (IPE), inner pattern evolving in the pattern taxonomy model. The summarized results are described in Fig. 1. Since not all methods can complete all tasks in the last 50 TREC topics. As aforementioned, itemset-based data mining methods struggle in some topics as too many candidates are generated to be processed. In addition, results obtained based on the first 50 TREC topics are more practical and reliable since the judgment for these topics is manually made by domain experts, whereas the judgment for the last 50 TREC topics is created based on the metadata tagged in each document.

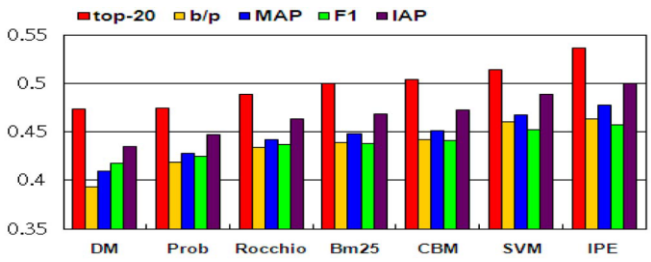

Fig 1: Comparison of PTM (IPE) and other major models in five measures for the 100 topics

IV. CONCLUSION

Text mining of repair verbatims for fault diagnosis of railway systems poses a big challenge due to unstructured verbatims, 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 $\chi 2$ 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. The proposed bi-level feature extraction method has been evaluated by RTP /RFP and F1-measure with a real data set collected by a railway company in China. The experiments show that the diagnosis results of the proposed feature fusion method, especially for minority fault classes, are much better than those of the traditional ones, such as $\chi 2$ statistics and information gain. Efficient feature fusion methods play an important role in feature extraction. Therefore, such powerful methods as parallel feature fusion should be further researched to improve the proposed method's performance Other merging learning methods should also be explored for better imbalanced classification.

REFERENCES

[1] 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.

[2] W. Wang, H. Xu, and X. Huang, "Implicit feature detection via aconstrained topic model and SVM," in Proc. Conf. Empirical Methods Natural Lang. Process., Seattle, WA, USA, 2013, pp. 903–907.

[3] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan, "Feature selection for high-dimensional imbalanced data," Neuro computing, vol. 105, pp. 3–11,Apr. 2013.

[4] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Constrained LDA for grouping productfeatures in opinion mining," in Proc. 15th Pacific-Asia Conf. Adv. Knowl.Discov. Data Mining, Shenzhen, China, 2011, vol. 1, pp. 448–459.

[5] X. Ding, Q. He, and N. Luo, "A fusion feature and its improvementbased on locality preserving projections for rolling element bearingfault classification," J. Sound Vibration, vol. 335, pp. 367–383,Jan. 2015.

[6] L. Huang and Y. L. Murphey, "Text mining with application to engineeringdiagnostics," in Proc. 19th Int. Conf. IEA/AIE, Annecy, France, 2006,pp. 1309–1317.

[7] J. Silmon and C. Roberts, "Improving switch reliability with innovativecondition monitoring techniques," Proc. IMechE, F C J. Rail RapidTransit, vol. 224, no. 4, pp. 293–302, 2010.

[8] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," J. Mach.Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.

[9] J. Chang, J. Boyd-Graber, C.Wang, S. Gerrish, and D. Blei, "Reading tealeaves: How humans interpret topic models," Neural Inf. Process. Syst.,vol. 22, pp. 288–296, 2009.

[10] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanceddata," in Proceedings of the 2008 European Conference on MachineLearning and Knowledge Discovery in Databases-Part I. Berlin,Germany: Springer-Verlag, 2008, pp. 241–256.

[11] T. Kailath, "The divergence and Bhattacharyya distance measures in signalselection," IEEE Trans. Commun. Technol., vol. 15, no. 1, pp. 52–60,Feb. 1967.

[12] J. Yang, J. Yang, D. Zhang, and J. Lu, "Feature fusion: Parallel strategyvs. serial strategy," Pattern Recognit., vol. 36, no. 6, pp. 1369–1381,Jun. 2003.

[13] C. Drummond and R. C. Holte, "C4. 5, class imbalanced, and cost sensitivity:Why under-sampling beats over-sampling," in Proc. WorkshopLearn. Imbalanced Datasets II, ICML, Washington, DC, USA, 2003,pp. 1–8.

[14] X. Liu, J. Wu, and Z. Zhou, "Exploratory undersampling for classimbalancelearning," IEEE Trans. Syst., Man Cybern., B, vol. 39, no. 2,pp. 539–550, Apr. 2009.

[15] D. Margineantu and T. G. Dietterich, "Learning decision trees for lossminimization in multi-class problems," Dept. Comput. Sci., Oregon StateUniv., Corvallis, OR, USA, Tech. Rep., 1999.

[16] M. V. Joshi, R. Agarwal, and V. Kumar, "Predicting rare classes: Canboosting make any week learner strong?" in Proc. 8th ACM SIGKDDInt. Conf. Knowl. Discov. Data Mining, Edmonton, AB, Canada, 2002,pp. 297–306.

[17] Y. Tang, Y. Zhang, and N. V. Chawla, "SVMs modeling for highly imbalancedclassification," IEEE Trans. Syst., Man Cybern., B, vol. 39, no. 1,pp. 281–288, Feb. 2009.

[18] G. Weiss, "Mining with rarity: A unifying framework," ACM SIGKDDExplorations Newslett.—Spec. Issue Learn. Imbalanced Datasets, vol. 6,no. 1, pp. 7–19, Jun. 2004.

[19] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced classdistribution and naive Bayes," in Proc. 16th Int. Conf. Mach. Learn., Bled,Slovenia, 1999, pp. 258–267.

[20] Z. Zheng, X.Wu, and R. Srihari, "Feature selection for text categorizationon imbalanced data," ACM SIGKDD Explorations Newslett.—Spec. IssueLearn. Imbalanced Datasets, vol. 6, no. 1, pp. 80–89, Jun. 2004.

