SCIENCE CHINA
Information Sciences

• LETTER •

July 2024, Vol. 67, Iss. 7, 179301:1–179301:2
https://doi.org/10.1007/s11432-023-4055-1

An ensemble and cost-sensitive learning-based root cause diagnosis scheme for wireless networks with spatially imbalanced user data distribution

Qi WANG[1], Zhiwen PAN[1,2]* & Nan LIU[1]

[1] *National Mobile Communications Research Laboratory, Southeast University, Nanjing* 210096*, China;*
[2] *Purple Mountain Laboratories, Nanjing* 211100*, China*

Received 29 November 2023/Revised 12 March 2024/Accepted 31 May 2024/Published online 25 June 2024

Root cause diagnosis, a key component of self-healing, plays a vital role in fault management. The spatially imbalanced network key performance indicators (KPIs) reported by users increase the difficulty of identifying root causes. In [1,2], the image inpainting technique is inspired to address issues caused by sparse reports across the coverage area. However, the assumption that reports are sparse throughout entire coverage areas is impractical, as it is rare for reports to be uniformly unavailable. Moreover, existing data quality improvement technologies, such as image inpainting, introduce significant time consumption in data generation and processing. Additionally, cost-sensitive learning methods are seldom used. In [3], a cost-sensitive learning method calculates the rescale ratio to address the imbalance of positive and negative samples. The algorithm in [4] includes severity level information but does not improve root cause diagnosis performance. Furthermore, labeling feature vectors with root causes is labor-intensive, and labeling severity levels is particularly challenging, as discussed in [5]. Motivated by these challenges, we propose a novel feature extraction method based on data similarity to address problems caused by imbalanced user data distribution. To further improve performance, we employ a cost-sensitive support vector machine (SVM) that allocates different misclassification costs to different severity levels of faults and a severity level evaluation algorithm to assist its implementation.

System model. Several base stations and user equipment (UE) are considered in this analysis. To simplify the analysis, UE includes a single antenna, while each base station is equipped with three uniform rectangular arrays. Pieces of UE are distributed unevenly in the horizontal plane. Each UE is serviced by its nearest base station, with other cells causing interference. Pieces of UE conduct minimization of drive test (MDT) reports. As discussed in [4], three root causes are considered: small antenna tilt resulting in excessive uptilt (EU), large antenna tilt leading to excessive downtilt (ED), and a significant drop in the transmit power of base stations causing excessive reduced power (ERP). Network cause classification is assessed using network KPIs, including signal-to-interference-plus-noise ratios (SINR) and reference signal received power (RSRP).

Evaluation of severity level based on Mahalanobis distance. Three severity levels are taken into account: serious, medium, and slight. Assume that the training set is partly labeled with severity levels, and KPIs in normal networks are available. The proposed algorithm aims to label the remaining training samples. The Mahalanobis distance between sample KPIs and KPIs in normal networks serves as the indicator for severity level labeling. For $\boldsymbol{p} = (p_1, p_2, \ldots, p_l)^{\mathrm{T}}$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_l)^{\mathrm{T}}$, this distance is defined as $d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{(\boldsymbol{p}-\boldsymbol{q})^{\mathrm{T}} \cdot \mathrm{cov}(\boldsymbol{p}, \boldsymbol{q}) \cdot (\boldsymbol{p}-\boldsymbol{q})}$, where $\mathrm{cov}(\boldsymbol{p}, \boldsymbol{q}) = E(\boldsymbol{p}^{\mathrm{T}}\boldsymbol{q}) - E(\boldsymbol{p}) \cdot E(\boldsymbol{q})$, where $E(\cdot)$ represents the mean. Define M as the number of samples in the training set, and the mth sequence of network KPIs is $\boldsymbol{x}_m = [\mathrm{RSRP}_1,\ \mathrm{SINR}_1, \ldots,\ \mathrm{RSRP}_i,\ \mathrm{SINR}_i, \ldots,\ \mathrm{RSRP}_I,\ \mathrm{SINR}_I]$, where RSRP_i and SINR_i are the RSRP and SINR of the ith user in the coverage area, respectively, I is the number of users, $m = 1, 2, \ldots, M$, and $i = 1, 2, \ldots, I$. M_n network KPIs under normal conditions are collected, and the m_nth network KPIs under normal conditions are $\boldsymbol{x}\_n = [\mathrm{RSRP}_{1\_n},\ \mathrm{SINR}_{1\_n}, \ldots,\ \mathrm{RSRP}_{i\_n},\ \mathrm{SINR}_{i\_n}, \ldots,\ \mathrm{RSRP}_{I\_n},\ \mathrm{SINR}_{I\_n}]$, where RSRP_{i_n} and SINR_{i_n} are the RSRP and SINR of the ith user, and $m\_n = 1, 2, \ldots, M\_n$. M_k samples are labeled with known severity level information, forming the set $\boldsymbol{I}_k = [i_k^1, \ldots, i_k^{M_k}]$. The set of those KPIs with corresponding labels is $\boldsymbol{X}\_\mathrm{label} = \{(\boldsymbol{x}_{i_k^1}, y_{s\_i_k^1}), \ldots, (\boldsymbol{x}_{i_k^{M_k}}, y_{s\_i_k^{M_k}})\}$, where $y_{s\_i_k^1}, \ldots, y_{s\_i_k^{M_k}} \in \{1, 2, 3\}$. If the corresponding severity level is serious, $y_{s\_i_k^{m_k}} = 1$. When a fault is labeled with a medium severity level, $y_{s\_i_k^{m_k}} = 2$. Similarly, $y_{s\_i_k^{m_k}} = 3$ is set for a fault with a slight severity level. $m_k = 1, \ldots, M_k$. To deal with the challenge caused by the imbalanced distribution of user data, the whole coverage area containing N_c cells is separated into $N_s = 2N_c$ regions. In each region, only a small amount of user data are available with I_s users in the sth region, $s \in \{1, \ldots, N_s\}$. The KPIs of those users in sample networks and normal networks are separately described as $\boldsymbol{x}_m^s$ and $\boldsymbol{x}_{m\_n}^s$. The indicator of the severity level prediction algorithm based on distance is defined as follows:

$$\mathrm{dis}_m = \max_s \sum_{m\_n=1}^{m\_n=M\_n} d(\boldsymbol{x}_m^s, \boldsymbol{x}_{m\_n}^s). \tag{1}$$

$\boldsymbol{N}_{11}$, $\boldsymbol{N}_{12}$, and $\boldsymbol{N}_{13}$ are the sets of samples with serious,

* Corresponding author (email: pzw@seu.edu.cn)

medium, and slight labels. $\mathrm{th}_1 = \frac{1}{2}(\min_{m\in \boldsymbol{N}_{11}} \mathrm{dis}_m + \max_{m\in \boldsymbol{N}_{12}} \mathrm{dis}_m)$ aims to distinguish serious and medium severity levels while $\mathrm{th}_2 = \frac{1}{2}(\min_{m\in \boldsymbol{N}_{12}} \mathrm{dis}_m + \max_{m\in \boldsymbol{N}_{13}} \mathrm{dis}_m)$ differentiates between medium and slight severity levels. For the sample KPIs, calculate the indicator dis_n using (1). If $\mathrm{dis}_n \geqslant \mathrm{th}_1$, identify it as a serious fault. If $\mathrm{th}_2 \leqslant \mathrm{dis}_n < \mathrm{th}_1$, classify it as a medium fault. Otherwise, it is a slight fault.

Feature extraction. Feature extraction transfers the original features to a new feature space through mapping, with the purpose of filtering out useful information and reducing redundancy. There is no doubt that the greater the indicator in the severity level prediction in (1), the less the similarity between the sample KPIs and the normal-condition KPIs. This results in more representative fault characteristics and a greater impact on the root cause diagnosis results. In the proposed feature extraction algorithm, three sections corresponding to the three greatest similarity indicators are considered. The section corresponding to the maximum value of similarity indicators is defined as s_1, while the sections corresponding to the second and third maximum values are defined as s_2 and s_3. Randomly select N_1, N_2, and N_3 users in s_1, s_2, and s_3, respectively, where $N_1 \geqslant N_2 \geqslant N_3$. For $\boldsymbol{x}_m$, the KPIs of those selected users make up the extracted feature vector defined as $\boldsymbol{x}_{m\_1}$. The training set is $\boldsymbol{X}_{MD} = \{(\boldsymbol{x}_{1\_1}, y_1), \ldots, (\boldsymbol{x}_{M\_1}, y_M)\}$, where $y_m \in \{1, 2, 3\}$. If $\boldsymbol{x}_m$ is the KPIs vector of an EU fault, $y_m = 1$. If $\boldsymbol{x}_m$ is the KPIs vector of an ED fault, $y_m = 2$. Otherwise, $y_m = 3$. Moreover, principal component analysis (PCA) is applied as another feature extraction method, mapping the mth feature vector $\boldsymbol{x}_m$ to $\boldsymbol{x}_{m\_2}$. Thus, the training set is built as $\boldsymbol{X}_{\mathrm{PCA}} = \{(\boldsymbol{x}_{1\_2}, y_1), \ldots, (\boldsymbol{x}_{M\_2}, y_M)\}$.

Cost-sensitive SVM-based root cause diagnosis. After feature extraction, $\boldsymbol{X}_{MD}$ and $\boldsymbol{X}_{\mathrm{PCA}}$ are obtained as inputs to train classifiers independently. Owing to the similarity in the training phase, $\boldsymbol{X}_{MD}$ is utilized as an example to illustrate the details and we only need to replace $\boldsymbol{X}_{MD}$ with $\boldsymbol{X}_{\mathrm{PCA}}$ to realize the training phase of classifiers based on $\boldsymbol{X}_{\mathrm{PCA}}$. Cost learning is used in the prediction phase of SVM. Define the cost matrix $\mathbf{Cost} = [\mathbf{Cost}(k, j)]$ as shown in Table 1, where $\mathbf{Cost}(k, j)$ indicates the cost of class k being classified as class j, where $i, j \in \{1, \ldots, 9\}$. $i, j = 1$ indicates serious EU class, $i, j = 2$ medium EU class, $i, j = 3$ slight EU class, $i, j = 4$ serious ED class, $i, j = 5$ medium ED class, $i, j = 6$ slight ED class, $i, j = 7$ serious ERP class, $i, j = 8$ medium ERP class, and $i, j = 9$ slight ERP class. The EU class with serious severity is defined as **EU1**, the EU class with medium severity is defined as **EU2**, the ED class with medium severity is defined as **EU3**, and the same representations are defined for ED and ERP classes. c_1 denotes the misclassification costs for faults with a serious severity level while the misclassification costs for faults with a medium severity level and slight severity level are defined as c_2 and c_3, where $c_1 \geqslant c_2 \geqslant c_3$.

Ensemble learning. During the creation of $\boldsymbol{X}_{MD}$, there is strong randomness in the selection of users and $\boldsymbol{X}_{MD}$ treats the challenge caused by the imbalanced user data distribution. Thus, the output of the trained cost-sensitive SVM performs well with imbalanced data but can sometimes be unstable. $\boldsymbol{X}_{\mathrm{PCA}}$ is a deterministic training set that does not involve special treatment for imbalanced user data distribution. As a result, the diagnosis results of the cost-sensitive SVM trained with $\boldsymbol{X}_{\mathrm{PCA}}$ are stable but ordinary. Moreover, the inspiration for generating $\boldsymbol{X}_{MD}$ is to handle the challenge of imbalanced user data distribution, which is reflected in favorable simulation results. In this regard, ensemble learning plays a major role in upgrading performance. The classifier based on ensemble learning will identify sample network KPIs $\boldsymbol{x}_n$ as y_{p_n} if and only if

Table 1 Cost matrix

	EU1	**EU2**	**EU3**	**ED1**	**ED2**	**ED3**	**ERP1**	**ERP2**	**ERP3**
EU1	0	0	0	c_1	c_1	c_1	c_1	c_1	c_1
EU2	0	0	0	c_2	c_2	c_2	c_2	c_2	c_2
EU3	0	0	0	c_3	c_3	c_3	c_3	c_3	c_3
ED1	c_1	c_1	c_1	0	0	0	c_1	c_1	c_1
ED2	c_2	c_2	c_2	0	0	0	c_2	c_2	c_2
ED3	c_3	c_3	c_3	0	0	0	c_3	c_3	c_3
ERP1	c_1	c_1	c_1	c_1	c_1	c_1	0	0	0
ERP2	c_2	c_2	c_2	c_2	c_2	c_2	0	0	0
ERP3	c_3	c_3	c_3	c_3	c_3	c_3	0	0	0

$$y_{p\_n} = \arg\min_{k} \sum_{j=1}^{9} [P(\text{predicted class}{=}j|\boldsymbol{x}_{n\_1}) \cdot \mathbf{Cost}(k, j) + a \cdot P(\text{predicted class}{=}j|\boldsymbol{x}_{n\_2}) \cdot \mathbf{Cost}(k, j)], \tag{2}$$

where $P(\text{predicted class}{=}j|\boldsymbol{x}_{n\_1})$, $P(\text{predicted class}{=}j|\boldsymbol{x}_{n\_2})$ denote the probability that the classification output is j for $\boldsymbol{x}_{n\_1}$, $\boldsymbol{x}_{n\_2}$; $\boldsymbol{x}_{n\_1}$, $\boldsymbol{x}_{n\_2}$ are the outputs of $\boldsymbol{x}_n$ via two feature extraction methods; and a is a weighting parameter.

Experiments and results. The proposed algorithm outperforms the other five algorithms. The detailed results are presented in Appendix A.

Conclusion. A novel feature extraction method that can tolerate imbalanced user data is proposed. A cost-sensitive SVM assigns different misclassification costs to faults with different severity levels to optimize the cause diagnosis process. The simulation results demonstrate the effectiveness and superiority of the proposed algorithm.

Acknowledgements This work was partially supported by National Key Research and Development Project (Grant No. 2020YFB1806805) and Qualcomm. We would like to specifically thank Qualcomm engineer Dr. Changlong XU for his support and encouragement.

Supporting information Appendix A. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Riaz M S, Qureshi H N, Masood U, et al. A hybrid deep learning-based (HYDRA) framework for multifault diagnosis using sparse MDT reports. IEEE Access, 2022, 10: 67140–67151

2 Park J, Kim B, Kim J, et al. Heterogeneous phased array architecture consisting of AoD and AiP to enhance spherical beamforming coverage for 5G/6G cellular handsets. In: Proceedings of the 15th European Conference on Antennas and Propagation (EuCAP), 2021. 1–4

3 Wang Y, Zhu K, Sun M, et al. An ensemble learning approach for fault diagnosis in self-organizing heterogeneous networks. IEEE Access, 2019, 7: 125662–125675

4 Chen K-F, Lin C-H, Lee M-C, et al. Deep learning-based multi-fault diagnosis for self-organizing networks. In: Proceedings of IEEE International Conference on Communications, 2021. 1–6

5 Chen M, Zhu K, Wang R, et al. Active learning-based fault diagnosis in self-organizing cellular networks. IEEE Commun Lett, 2020, 24: 1734–1737