Fuzzy Output Support Vector Machine Based Incident Ticket Classification

SUMMARY Incident ticket classification plays an important role in the complex system maintenance. However, low classification accuracy will result in high maintenance costs. To solve this issue, this paper proposes a fuzzy output support vector machine (FOSVM) based incident ticket classification approach, which can be implemented in the context of both two-class SVMs and multi-class SVMs such as one-versus-one and one-versus-rest. Our purpose is to solve the unclassifiable regions of multi-class SVMs to output reliable and robust results by more fine-grained analysis. Experiments on both benchmark data sets and real-world ticket data demonstrate that our method has better performance than commonly used multi-class SVM and fuzzy SVM methods.

key words: incident ticket classification, support vector machine, multi-class classification, fuzzy membership function

1. Introduction

For the maintenance and management of a complex IT infrastructure, when an event which is not part of the standard system operation and which may cause an interruption or a reduction happens is captured by the monitoring system, an incident ticket manifesting the event is automatically created. In general, a ticket contains some unstructured texts describing problem symptoms using the natural language. Once an incident ticket appears, we need rapid allocation of skilled maintenance experts to bring an abnormal service back to normal, which highly depends on the accurate ticket classification. As an initial step of rapid management, ticket classification is used to identify problem types of tickets based on the problem descriptions recorded in tickets [1], [2]. In a typical ticket system, the ticket classification is done manually by system administrators to assign a problem type. This manual process is time-consuming and error-prone. Hence, we need an automated approach to classify incident tickets with a high accuracy.

Since ticket problem descriptions are made of unstructured text and support vector machines (SVMs) have been successfully applied to many text classification scenes including email, news, and so on, we use it to solve the problem of incident ticket classification. However, the existing multi-class SVMs can not perfectly resolve the problem of the unclassifiable regions. When they are applied to ticket classification, they will suffer from a relatively high misclassification cost. Thus, this paper proposes a decision margin based fuzzy output SVM approach to reduce the unclassifiable regions and improve the accuracy of the incident ticket classification.

The rest of the paper is organized as follows: Sect. 2 gives related work, followed by primary concepts regarding fuzzy support vector machines in Sect. 3. Section 4 presents a fuzzy output SVM approach based on decision margin for the multi-class SVM techniques. Section 5 validates it on several benchmark data sets and a real-world ticket dataset; and finally, Sect. 6 concludes the paper.

2. Related Work

The incident ticket classification problem belongs to document classification in nature, where a document is a short free-text ticket problem description. Since we discuss a ticket classification approach based on SVMs, which is a supervised approach, we review the literature of SVMs and supervised ticket classification approaches.

2.1 Supervised Ticket Classification

If problem type information in historical tickets is available in the training ticket data, supervised ticket classification algorithms are also applicable. The most popular approach is to apply machine-learning techniques to automatically build a classifier on a set of pre-classified tickets to classify new tickets. Various supervised machine-learning techniques proposed for automatic text classification, such as support vector machines (SVM), Naïve Bayes, and maximum entropy, have been applied for maintenance and incident ticket classification [2]–[6]. For instance, [4] applies an SVM to predict the most appropriate ticket resolution group. [3] proposes a Multinomial Naïve Bayes (MNB) algorithm to classify tickets. In summary, because of the size and complexity of incident tickets, supervised classification algorithms have been the method of choice for incident ticket classification, relying on labeled tickets from a managed infrastructure to automatically create signatures for similar infrastructure.

2.2 SVMs

SVMs gain wide application due to its high generalization
ability and better performance than other traditional learning machines in recent years. But SVMs still suffer from the problem of unclassifiable regions [7], where the samples are non-determinable. Traditional Multi-class SVMs classify unclassifiable samples into a class randomly. It will decrease the generalization of learning machines. Moreover, in some practical problems such as in medical diagnosis and ticket classification, these unclassifiable samples draw more attention and should be given more precise analysis. There are mainly two ways of solving unclassifiable samples: one is to use continuous decision functions, and the other is to apply a fuzzy support vector machines (FSVM) [8]–[10].

The core study of FSVM is how to produce the fuzzy membership function. Some studies adopt a fuzzy membership function to each input point and reformulate the SVM so that different input points can make different contributions to the learning of decision surface [11]. Using the decision function obtained by training the SVM, [12] defined a truncated polyhedral pyramidal membership function for each class to solve the unclassifiable regions. In theory, the generalization ability of their proposed FSVM is superior to that of conventional SVMs. Many FSVM models [13] can be considered as the modifications or extensions of SVMs to reduce the effect of outliers or noises in training samples and have successfully applied to many applications such as text classification, fault diagnosis, and so on. Existing FSVM can give a reasonable decision for both classifiable and unclassifiable samples, but there are still two problems that need to be considered: (1) a sample is unclassifiable if the membership function gets the same maximum on two or more classes; (2) the classification results are unreliable if the difference of the values of decision function on different classes is very small.

3. Fuzzy Support Vector Machine

Consider training data pairs: where belongs to one of the classes, for a two-class classification problem, while for a k-class (k > 2) classification problem. In the basic form, SVM learns linear decision rules described by a weight vector and a threshold value. Its basic idea is to map data into a high dimensional space and find a separating hyperplane with the maximal margin. Thus, the linear two-class SVM model is:

$$\min_{w,b} \frac{1}{2} w^T w$$
$$s.t \quad y_i (w \cdot x_i + b) \geq 1, \ 1 \leq i \leq l$$

(1)

However, in real-world applications, even the linearly separable data may not be ideally linearly separable because of measurement errors or noises. Thus, we relax the constraints of Eq. (1) slightly to allow misclassified points by introducing a slack variable $|\xi_i| \geq 0$. In this soft margin SVM, data points on the incorrect side of the margin boundary have a penalty that increases with the distance from it. The number of misclassifications is reduced by solving:

$$\min \begin{cases} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\
\end{cases}$$
$$s.t \quad y_i (w^T \cdot x_i + b) \geq 1 - \xi_i \quad and \quad \xi_i \geq 0, \ i = 1, \ldots, l$$

(2)

where the parameter $C$ controls the trade-off between the slack variable penalty and size of the margin. Decreasing the value of $C$ means to allow more misclassifications and to relax the hyperplane tension. To solve this constrained quadratic optimization problem, we find its dual using Lagrange multipliers and its dual form is:

$$\max \begin{cases} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\
\end{cases}$$
$$s.t \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \quad and \quad 0 \leq \alpha_i \leq C, \ i = 1, \ldots, l$$

(3)

The above SVM is an essentially linear model, but it can be easily generalized to non-linear decision rules by replacing the inner-product $(x_i \cdot x_j)$ with a kernel function $k(x_i, x_j) = \langle \phi(x_i) \cdot \phi(x_j) \rangle$, where the function $\phi$ is a real one that projects the data into a higher dimensional space. We only need to find the kernel matrix whose entry $(i, j)$ will correspond to the inner product of feature space vectors $\langle \phi(x_i) \cdot \phi(x_j) \rangle$ without worrying about the images of $\phi(x_i)$ and $\phi(x_j)$.

In general, a multi-class problem is converted into a certain number of two-class problems. There are two commonly used strategies: one versus rest (1-v-r) and one versus one (1-v-1). In the 1-v-r method, the $i$th SVM classifier is trained by assigning positive labels to all the samples in the $i$th class, and negative labels to all other samples. The training time of the standard method measures linearly with $k$. In the 1-v-1 method, it constructs $k(k - 1)/2$ binary classifiers by training on only two out of classes each time. Thus, the size of the 1-v-1 classifiers may grow super-linearly with $k$. The combination of these binary classifiers to determine the label assigned to each new input can be made by different policies. For example, we count the votes of each sub-classifier using the majority vote policy and the class with most votes is the final decision. There will remain some unclassifiable regions while converting a multi-class SVM into a two-class SVM with the majority vote algorithm (see Fig. 1).

To resolve the unclassifiable regions for 1-v-r and 1-v-1
strategy, fuzzy SVM is an effective way [8]. Take 1-v-1 as an example, which can be briefly introduced as follows. Let the $i$th decision function that classifies class $i$ and class $j$ be

$$f_{ij}(x) = (w_{ij} \cdot x) + b_{ij}$$

(4)

where $f_{ij}(x) = -f_{ji}(x)$. For an input vector $x$, we calculate

$$f_i(x) = \sum_{j \neq i, j=1}^{n} sgn(f_{ij}(x))$$

(5)

where $sgn = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$ and classify using the majority vote policy.

$$label(x) \equiv \arg \max_{j=1..k} f_j(x)$$

(6)

In this formulation, however, there are unclassifiable region remaining (as shown in Fig. 1, the shadow regions), where each $f_i(x)$ has the same value. [12] introduced two new membership functions:

$$DM(x) = \max_{j=1..k} \left( f_{ij}(x) - \max_{j=1..k} f_j(x) \right)$$

(7)

and

$$m_i(x) = \frac{1}{k-1} \sum_{j \neq i, j=1}^{k} m_{ij}(x)$$

(8)

where $m_{ij}(x) = \begin{cases} 1 & f_{ij}(x) \geq 1 \\ f_{ij}(x) & -1 < f_{ij}(x) < 1 \\ 0 & f_{ij}(x) \leq -1 \end{cases}$ is classified into the class.

$$label(x) \equiv \arg \max_{j=1..k} m_i(x)$$

(9)

Thus, the unclassified region is shown in Fig. 1 (1-v-1) is resolved as shown in Fig. 2.

4. FOSVM

Although the existing FSVM methods can give a proper decision for unclassifiable samples, they have some limitations. On one hand, the fuzzy membership functions proposed in these algorithms may acquire the same maximum on two or more classes, thus the samples are still unclassifiable. Meanwhile, the difference of decision function value on different classes may be small in unclassifiable regions. In these cases, the classification results are unreliable and sensitive to the noise. On the other hand, in some application domains, such as in medical diagnosis and ticket classification, these unclassifiable samples are typically key points of analysis. They need be given more robust and credible classification results.

To deal with the abovementioned issues, we propose a decision-margin based fuzzy output support vector machines (FOSVM) framework which can be implemented in the context of both two-class SVM and multi-class SVM such as 1-v-1 and 1-v-r.

4.1 One-versus-Rest FOSVMs

The 1-v-r method resolves the following k two-class problems:

$$\min_{w, b_i, \xi_i} \frac{1}{2} \|w^j\|^2 + C \sum_{i=1}^{l} \xi_i$$

$$\begin{align*}
(w^j \cdot x_i) + b^j - 1 + \xi_i & \geq 0, \quad \text{if } y_i = j \\
(w^j \cdot x_i) + b^j - 1 + \xi_i & \leq 0, \quad \text{if } y_i \neq j
\end{align*}$$

where $\xi_i \geq 0, i = 1 \ldots l$.

Then we can get k decision functions $f_j(x)$. Instead of the majority vote algorithm, we definite a new decisions function.

**Definition 1 (decision margin):** The difference between the maximal value and the second maximal value of the decision function is called decision margin (DM).

$$DM(x) = \max_{j=1..k} f_j(x) - \max_{j=1..k} f_j(x)$$

(11)

where $\max_{j=1..k} f_j(x)$ and $\max_{j=1..k} f_j(x)$ represent the maximum and second maximum of the decision function respectively.

We adopt the membership function as follows.

$$\mu(x) = \frac{1}{1 + e^{-\alpha DM(x) - \beta}}$$

(12)

where $DM(x)$ is decision margin defined as Eq. (11), $\alpha$ and $\beta$ are two adjustable parameters. The final output of classifier is:

$$label(x) = \begin{cases} \arg \max_{j=1..k} f_j(x) & \text{if } \mu(x) > \delta \\
\text{refuse} & \text{if } \mu(x) \leq \delta
\end{cases}$$

(13)

where $\delta$ is a threshold according to the given problem. If the decision margin is less than $\delta$, FOSVM refuses to predict.
the clear label.

For simplicity, take a three-class classification problem for an example, the whole feature space will be divided into three regions, as shown in Fig. 3.

- Clear Classifiable region: \( \max_{j=1,...,k} f_j(x) - \max_{j=1,...,k} f_j(x) \geq 2; \)
- Classifiable region: \( \delta < \max_{j=1,...,k} f_j(x) - \max_{j=1,...,k} f_j(x) < 2; \)
- Unclassifiable region: \( \max_{j=1,...,k} f_j(x) - \max_{j=1,...,k} f_j(x) \leq \delta, \delta < 2; \)

Membership function defined in Eq. (12) has the following character: it has the highest confidence level in clear classifiable regions, the lowest confidence level in unclassifiable regions, and the medium confidence level in classifiable regions. For example, if \( \delta = 0.5, \) the membership function has the following form as shown in Fig. 4, where \( \alpha = -\ln 0.25/(2 + \delta), \beta = \delta. \)

4.2 One-versus-One FOSVMs

Similar to the above 1-v-r FOSVM method, we can also construct the 1-v-1 FOSVM. First, we compute the decision value of class \( j \) by constructing \( k(k-1)/2 \) classification functions \( f_{ji}(x) \).

\[
 f_j(x) = \min_{i \neq j, i=1,...,k} f_{ji}(x) 
\]  

(14)

Then we compute the decision margin \( DM(x) \) according to Eq. (11). Finally, we get the output of classification via Eq. (12) and Eq. (13).

Our proposed method is also suitable for the two-class problem. In this case, \( DM(x) = f_i(x), i = 1, 2 \) and \( f_1(x) = -f_2(x) \). The values of parameters are \( \alpha = -\ln 0.25 \) and \( \beta = 0 \) respectively. Using the same decision function, we can get the final result. The membership function has the form as shown in Fig. 5. For example, if a sample belongs to one class with a possibility of 80%, it belongs to another class with 20% possibility.

To sum up, our proposed methods are defined so that, for the data in classifiable regions, the classification results are the same with the standard 1-v-1 and 1-v-r methods; for the data in partial classifiable regions, the classifier can also give rational results; for the data with the membership value less than the given threshold, the classifier will not give a hard label. In this case, we can adopt other methods to deal with these samples such as applying prior knowledge or nesting algorithm for multi-classification problems.

5. Experiments

In this section, we evaluate our methods on several benchmark data sets and the ticket dataset and compare them with the standard multi-SVM and popular FSVM.

5.1 Benchmark Datasets

Table 1 shows the summary of the datasets used in the experiments. These datasets are from UCI repository with the number of classes ranging from 2 to 6. Glass6 is an original glass dataset with 214 samples, glass3 is a subset of the glass dataset of three classes including ‘float processed building windows’, ‘non-float processed building windows’ and ‘non-window glass’, totally 197 samples.

We tried different kernels including linear, polynomial, quadratic, Gaussian and RBF, and parameters \( C \) and \( \delta \) were determined by the ten-fold cross-validation method. Their
Table 2  Performance compare on four benchmark datasets

| Data sets | Ratio of unclassifiable samples (%) | 1-v-1 | 1-v-r | FSVM (1-v-r) | FOSVM (1-v-r) |
|-----------|-------------------------------------|-------|-------|--------------|--------------|
|           |                                     | Acc ± var (%) | Acc ± var (%) | Acc ± var (%) | Acc ± var (%) |
| balance   | 6.42                                | 89.30 ± 2.02  | 90.37 ± 2.16  | 92.10 ± 2.08  | 92.51 ± 1.88  |
| breast    | 0                                   | 98.08 ± 0.0   | 98.08 ± 0.0   | 98.08 ± 0.0   | 98.08 ± 0.0   |
| glass3    | 10.17                               | 75.34 ± 2.13  | 74.58 ± 3.61  | 75.97 ± 2.54  | 76.27 ± 2.01  |
| glass6    | 13.85                               | 63.58 ± 2.56  | 61.53 ± 3.07  | 62.78 ± 2.17  | 63.86 ± 2.14  |

Table 3  Performance comparison on the ticket dataset

| No. | Algorithms                  | Ratio of unclassifiable samples (%) | Accuracy |
|-----|-----------------------------|-------------------------------------|----------|
|     |                             |                                     | precision | recall | F1 score |
| 1   | 1-v-1 SVM(tf-idf)           | 6.78                                | 0.8701    | 0.8192 | 0.8439   |
| 2   | 1-v-1 SVM(word2vec)         | 6.62                                | 0.8892    | 0.8199 | 0.8531   |
| 3   | 1-v-r SVM (tf-idf)          | 7.03                                | 0.8623    | 0.8091 | 0.8348   |
| 4   | 1-v-r SVM (word2vec )       | 6.95                                | 0.8734    | 0.8152 | 0.8433   |
| 5   | 1-v-r FSVM (tf-idf)         | 4.34                                | 0.9354    | 0.8928 | 0.9136   |
| 6   | 1-v-r FSVM (word2vec )      | 4.25                                | 0.9389    | 0.9001 | 0.9191   |
| 7   | 1-v-r FOSVM (tf-idf)        | 2.51                                | 0.9641    | 0.9131 | 0.9379   |
| 8   | 1-v-r FOSVM (word2vec )     | 2.47                                | 0.9682    | 0.9176 | 0.9422   |

best results are reported. The samples with the membership score less than the given threshold were reclassified using the method [14]. Each algorithm was repeated 10 times. The experimental results are shown in Table 2.

From Table 2, we observe that compared with standard multi-class SVM techniques, 1-v-1 and 1-v-r, the proposed decision-margin based FOSVM classifies the samples in a more fine-grained way. The samples with a higher confidence have the same classification results with other two methods, while the samples with a lower confidence will be handled specially. Our method has better performance in terms of accuracy on the problems that require a high accuracy; especially those applications that emphasize the ‘hard samples’ such as in medical diagnosis and credit risk assessment.

5.2  Ticket Dataset

The real world ticket datasets are collected from an account of a large cloud-oriented IT service center. This account consists of over 200 monitored servers and network devices. The dataset has 100K+ tickets that cover a period of eight months. The number of problem types covered by these tickets is 95, and the distribution of tickets of different problem types is highly unbalanced. We choose 10000 tickets that belong to the largest K problem types according to their original proportions as the experimental dataset.

The critical ticket attributes used in our experiments include “problem type” and “description”, where the attribute “problem type” denotes the problem cause, and the attribute “description” denotes the problem symptom. Two ticket representation approaches including the tf-idf term weighted scheme and word2vec are applied to obtain a vector representation for SVM. We compare the FOSVM with 1-v-1 SVM, 1-v-r SVM and FSVM. We randomly select 75% samples as the training set and the rest 25% as the testing set. Other settings are the same as those used in Sect. 5.1.

The evaluation measures used in this experiment are precision, recall, and the F1 measure (F1 Score). These measures are standard accuracy metrics used in classification problems, and their definitions are expressed as follows:

\[
F1 \text{ score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}},
\]

\[
\text{precision} = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} TP_i + FP_i}, \quad \text{and}
\]

\[
\text{recall} = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} TP_i + FN_i},
\]

where K is the number of chosen problem types, \(TP_i\), \(FP_i\) and \(FN_i\) are True Positives (how many tickets were classified as a specified problem type \(c_i\) and they were indeed labeled as \(c_i\) in the data set), False Positives (how many were classified as a specified type \(c_i\) while they truly are with a type \(c_j\), \(c_j \neq c_i\)) and false negatives (how many were classified as a specified type while they truly are with a type \(c_i\), \(c_j \neq c_i\)), respectively.

Table 3 shows the performance comparison in terms of accuracy with other commonly used multi-class SVM techniques. We can see that the algorithm FOSVM performs best in terms of the ratio of unclassifiable samples, precision, recall and F1 score. FOSVM can reduce the ratio of unclassifiable samples and obtain more believable results than other algorithms. Further, compared to the tf-idf representation approach, the word2vec-based representation
gets a better accuracy in terms of precision, recall and F1 score, which means that the ticket presentation has a positive impact on its classification performance.

6. Conclusions

When a critical system exhibits an incident during its operation, system maintenance teams are expected to rapidly allocate skilled resources to bring an abnormal service back to normal as short as possible. In a typical ticket management system, the ticket classification is done manually by system administrators to assign a problem type, which is time-consuming and error-prone, especially when there are a large number of tickets. In this paper, the proposed a decision margin based FOSVM can be used to deal with this issue in an automated way. Our algorithm can resolve the unclassifiable regions of multi-class SVMs to improve the classification accuracy. Experiments on both the benchmark datasets and real-world ticket data have validated the effectiveness of the proposed algorithm.

Moreover, we observed that different incident ticket representation approaches has a significant impact on classification performance. We adopted two different representation approaches, including the traditional tf-idf and word2vec-based one, and found that the ticket representation using the word embedding technology can help us improve classification accuracy.

Acknowledgments

This research is supported by by Key Scientific and Research Project in University of Henan Province (15A520021).

References

[1] J. Xu, H. Zhang, W. Zhou, R. He, and T. Li, “Signature based trouble ticket classification,” Future Generation Computer Systems, vol.78, pp.41–58, 2017.
[2] J. Xu, L. Tang, and T. Li, “System situation ticket identification using svms ensemble,” Expert Systems with Applications, vol.60, pp.130–140, 2016.
[3] G. Son, V. Hazlewood, and G.D. Peterson, “On Automating XSEDE User Ticket Classification,” Conference on Extreme Science and Engineering Discovery Environment, pp.1–7, 2014.
[4] S. Agarwal, R. Sindghatta, and B. Sengupta, “SmartDispatch: enabling efficient ticket dispatch in an IT service environment,” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1393–1401, 2012.
[5] C. Kadar, D. Wiesmann, J. Iria, D. Husemann, and M. Lucic, “Automatic Classification of Change Requests for Improved IT Service Quality,” SrrI Global Conference, pp.430–439, 2011.
[6] J. Bogojeska, I. Giurgiu, D. Lanyi, G. Stark, and D. Wiesmann, “Impact of HW and OS type and currency on server availability derived from problem ticket analysis,” Network Operations and Management Symposium, pp.1–9, 2014.
[7] R. Li, A. Li, T. Wang, and L. Li, “Vector projection method for unclassifiable region of support vector machine,” Expert Systems with Applications, vol.38, no.1, pp.856–861, 2011.
[8] Y. Wu, L. Shen, and S. Zhang, “Fuzzy multiclass support vector machines for unbalanced data,” Control and Decision Conference, pp.2227–2231, 2017.
[9] R. Pruengkarn, K.W. Wong, and C.C. Fung, “Imbalanced data classification using complementary fuzzy support vector machine techniques and SMOTE,” IEEE International Conference on Systems, Man and Cybernetics, pp.978–983, 2017.
[10] J. Liu and E. Zio, “A scalable fuzzy support vector machine for fault detection in transportation systems,” Expert Systems with Applications, vol.102, pp.36–43, 2018.
[11] Y. Guerbai, Y. Chibani, and N. Abbas, “One-class versus bi-class SVM classifier for off-line signature verification,” International Conference on Multimedia Computing and Systems, pp.206–210, 2012.
[12] T. Inoue and S. Abe, “Fuzzy support vector machines for pattern classification,” International Joint Conference on Neural Networks, vol.2, pp.1449–1454, 2001.
[13] A.B. Ji, S. Chen, and Q. Hua, “Fuzzy classifier based on fuzzy support vector machine,” Journal of Intelligent & Fuzzy Systems, vol.26, no.1, pp.421–430, 2014.
[14] B. Liu, Z. Hao, and X. Yang, “Nesting algorithm for multi-classification problems,” Soft Computing, vol.11, no.4, pp.383–389, 2007.

Libo Yang received the B.S. and M.S. degrees in computer science from North China University of Water Resources and Electric Power, Zhengzhou, China in 2004 and 2011, respectively. He is currently pursuing the Ph.D. degree with North China University of Water Resources and Electric Power, Zhengzhou, China. His current research interests include machine learning, pattern recognition, data mining, and information management.