A Behavior Analysis Method Towards Product Quality Management

Congcong Ye[1], Chun Li[2], Guoqiang Li[1], Lihong Jiang[1(✉)], and Hongming Cai[1]

[1] Shanghai Jiao Tong University, Shanghai, China
{yecongcong,li.g,jianglh,hmcai}@sjtu.edu.cn
[2] China Shanghai Institute of Precision Measurement and Testing, Shanghai, China

Abstract. Quality management is the basic activity in industrial production. Assuring the authenticity of testing datasets is extremely important for the quality of products. Many visual tools or association analysis methods are used to judge the authenticity of testing data, but it could not precisely capture behavior pattern and time consuming. In this paper, we propose a complete framework to excavate the features of testing datasets and analyze the testing behavior. This framework uses min-max normalization method to pre-process datasets and optimized k-means algorithm to label the training datasets, then SVM algorithm is applied to verify the accuracy of our framework. Using this framework, we can get the features of dataset and homologous behavior model to distinguish the quality of datasets. Some experiments are carried to measure the complete framework and we use various visual formats to show these results and to verify our method.

Keywords: Behavior analysis · Product quality management · Data mining

1 Introduction

Product quality testing data are the key factors of industrial production to discover inconsistencies and other anomalies. Ensuring the quality of data has been a continuing concern to industrial production. Testing the quality of electronic components is a tough and humdrum job. Testers are always not willing to do this work, and they sometimes directly write the testing records based on their own experience or others' data. The feature of recording result depends on the testers' psycho-behavior. Different people have different strategies and it is extremely difficult to distinguish real testing data from fictional data.

At present, there is not much researches on quality of testing data, which tends to be regarded as real and authentic data. Some engineers only use simple data visual methods like histogram and scatter diagram to show the feature of data and judge the reliability of testing result. It is hard to distinguish the real data from fictional data because testers write testing records within the allowable range of error. Illusory testing data will decline the quality of industrial products and damage the corporate image.

To address the aforementioned challenges, a complete behavior analysis framework is proposed to distinguish the reliability of electronic components testing behavior. We

A. M. Tjoa et al. (Eds.): CONFENIS 2017, LNBIP 310, pp. 12–20, 2018.
https://doi.org/10.1007/978-3-319-94845-4_2

use an optimized k-means clustering method to find the aberrant data and give every record a label. After that one third of the dataset are used as training data to optimize the kernel function of support vector machine(SVM) and another dataset are used to test the accuracy of our method. This optimized SVM can be used to distinguish the real testing data from illusory data.

Our main contributions are summarized as follows:

- A joint probability distribution method is used to prove that the relationships of attributes are not fixed. The illusory testing data can not be distinguished by merely association analyzing.
- An optimized clustering algorithm is proposed to mine the inner features of testing datasets. This cluster algorithm is the optimized k-means algorithm based on density and distance.
- A complete behavior analysis framework is constructed to associate the dataset with behavior. The data model and behavior model can be changed according to different features of dataset.

The rest of this paper is organized as follows. Section 2 discusses the related work in this field. Section 3 presents the detailed behavior analysis process including data preprocessing, data mining algorithm and behavior analysis. Section 4 uses various methods to show the result of our methods and analyzes the experimental result. Section 5 concludes the work in this paper and illustrates the future direction of this work.

2 Related Work

Data mining is the analysis of datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [1]. It contains four major branches: clustering or classification, association rules and sequence analysis [8]. Especially, clustering algorithm is very useful when the number of classification is unknown. The major clustering algorithms are k-means, hierarchical clustering and self-organizing map algorithm. The initial K-means clustering algorithm aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [9]. The implementation of k-means is simple and the distance function can be changed based on the feature of datasets. Through K-means is the most common clustering algorithm, there are still some disadvantages about K-means like the choice of centers, the number of centers and so on. Many methods have been proposed to improve the performance of the K-means clustering algorithm. [16] combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters.

As for hierarchical clustering, it yields a dendrogram representing the nested grouping of patterns and similarity levels at which groupings change and the clustering process is performed by merging the most similar patterns in the cluster set to form a bigger one [3]. Hierarchical clustering uses greedy algorithm, so the result of hierarchical clustering is only local optimum. Another common clustering algorithm is self-organizing map, which is a neural network algorithm to create topographically ordered spatial

representations of an input data set using unsupervised learning [10]. All these common clustering algorithms can be chosen based on the features of datasets.

Another important branch of data mining is classification. Decision trees [5] is one of the most popular methods for classification [7]. In decision tree algorithm, each internal node split the instance space into two or more parts with the objective of optimizing the performance of classifier and every path from the root node to the leaf node forms a decision rule to determine which class a new instance belongs to [4].

In general, various clustering and classification algorithms can be used to dig out the inner features of datasets. However, there are not many researches about human behavior analysis. Data quality analysis method is not enough to find the behavior pattern which is reflected by the datasets if it only focuses on data features.

3 Methodology

3.1 A Framework of Mining Behavior Model from Datasets

The features of datasets are the reflection of testers' behavior. In order to analyze the relationship between the data model and behavior model, we construct a behavior analysis framework as Fig. 1. This framework consists of four modules: data pre-processing, training data construction module, behavior analysis module and data visualization module. In data pre-processing module, we can delete the outliers and use min-max

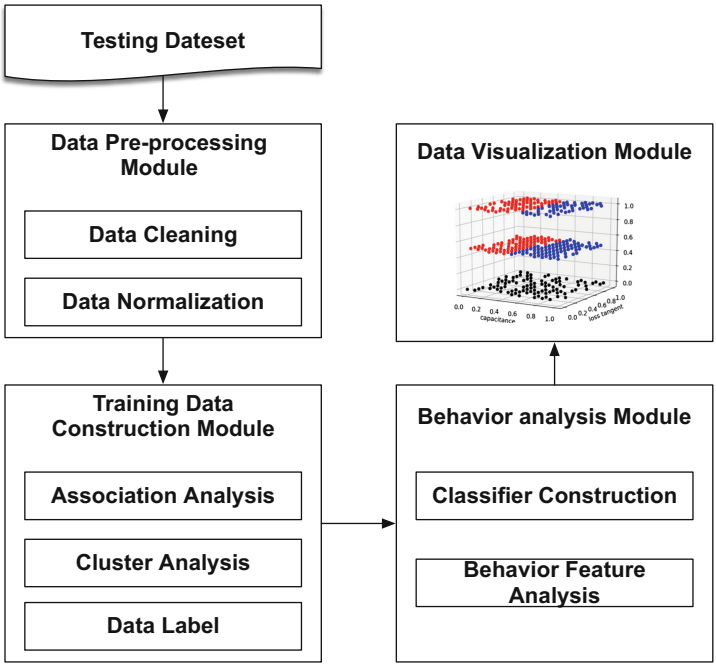


Fig. 1. A framework of mining behavior model from datasets

normalization method to process the datasets and improve the quality of datasets. In training data construction module, each dataset will be given a label through an optimized K-means algorithm and used to train the SVM classifier based on the association analysis and clustering result. At last, various visual methods are used to show the data mining results and behavior features.

Other data mining algorithms and behavior models, instead of k-means and SVM, can also be applied in case the datasets are replaced.

Data Pre-processing

Data pre-processing includes data preparation, compounded by integration, cleaning, normalization, transformation of data and data reduction tasks; such as feature selection, instance selection, discretization, etc. [6]. It will improve the quality of training data and increase reliability of data mining. First step of data pre-processing is data cleaning, which will dispose the noise and outliers. In order to ignore the effect of data dimension, the dataset will be normalized. There are many data normalization methods like min-max normalization, z-score normalization and decimal scaling normalization. The most common method is min-max normalization, which can transform every attribute into a value from zero to one. The calculation formula is defined as follow:

$$X^* = \frac{X - \min}{\max - \min} \tag{1}$$

After that we can use some attributes like mean value, variance, maximum, minimum and joint probability to excavate the features of data. All these features can reflect the fluctuation of every dataset and help engineers get an overview of dataset. However, only these methods are not enough for engineers to distinguish the authenticity of data and it needs more data mining methods to dig their inner features out.

Optimized Data Mining Algorithm

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems [11]. Clustering algorithms are usually used to process the datasets when their types of classification are not clear. In industrial production, the behaviors of testers depend on their different psychological activities. And we even don't know the number of different behavior types, which satisfies the conditions of clustering algorithms.

The most common clustering algorithm is k-means. Given a set of n points in d-dimension and the number of centers k, the problem is to determine k points in n as centers, so as to minimize the distance from each data point to its nearest center. The accuracy of k-means depends on the choice of k centers, which is also the main imperfection of k-means. In order to solve this problem and improve the accuracy of k-means, we optimize the original k-means algorithm and propose a k-means algorithm based on density and distance.

According to the relevant researches, the accuracy will be superior to others when the number of centers is set as the radication of the dataset number expect we have professional knowledge and ensure the clustering results. Center points are extremely crucial, but the original k-means algorithm chooses k centers randomly,

which is difficult to get the best cluster result. Then we optimize the strategy on choosing centers. Algorithm 1 represents the main framework of an optimized k-means algorithm based on density and distance. This algorithm has four parameters and the threshold values can be adjusted according to the features of datasets. This optimized method chooses the largest density point and other evenly distributed points as centers, which will avoid the situation that center points are excessive centralization. Using this framework, the noisy points or outliers will be ignored and the accuracy will be improved.

Algorithm 1: An optimized k-means algorithm based on density and distance

Input: Testing dataset dn, the number of clusters k, threshold density m, threshold radius r
Output: Clustering result S_k
1: **for** i = 0 to n **do**
2: calculate the distance of any two points in dataset dis(i;j) ;
3: **end for**
4: delete the points when their density is lower than threshold density;
5: Collect the points in high density area D;
6: find the largest density point as the first center, and put it in the center sets Z, then delete it from D;
7: find the furthest point from the first center and put it in the center sets Z, then delete it from D
repeat
8: calculate the distance product of every point in D with center sets;
9: find the furthest point and put it in Z;
10: delete it from the set D
11: **Until** the number of clusters do not reach k;
12: put every point in D to the suitable center which is most similar with it;
13: **return** Clustering result S_k

After clustering, we will give different clusters different labels and then use one third of these records about six hundred illusory and real records as training datasets to optimize the SVM classification. Finally, other datasets are used to judge the accuracy of our optimized algorithm and SVM. Not only this SVM can evaluate the accuracy of our labels, but also it can be used as an automated tool to distinguish illusory data quickly.

4 Experiments

The entire experiments were performed on a machine with 2.7 GHz Intel Core i5 cpu, equipped with 8G memory. The datasets come from an electronic component factory and they consist of four batches of product testing datasets. The dataset is defined as three dimensions, which are leakage current, minimum loss angle tangent and capacitance.

4.1 Association Analysis of Each Variate

Joint probability distribution [14] is an effective method to analyze the relationships between variates. In testing datasets, the crucial attributes are leakage current and min mum loss angle tangent. Hence we calculate their joint probability distribution and use 3-dimension figures to show the results.

Figure 2 contains four kinds of joint probability distributions on different batches testing datasets. Z-axis represents probability density and X or Y-axis are leakage current and minimum loss angle tangent. From these figures, we know the more points in front of the plane, the quality of datasets is higher. These four figures represent four types of the testing data: the multimodal state, the singlet state, the bowl state and the strip state. All these figures show the relationships between testing variates are not sure and it is not enough for us to distinguish the real datasets from others.

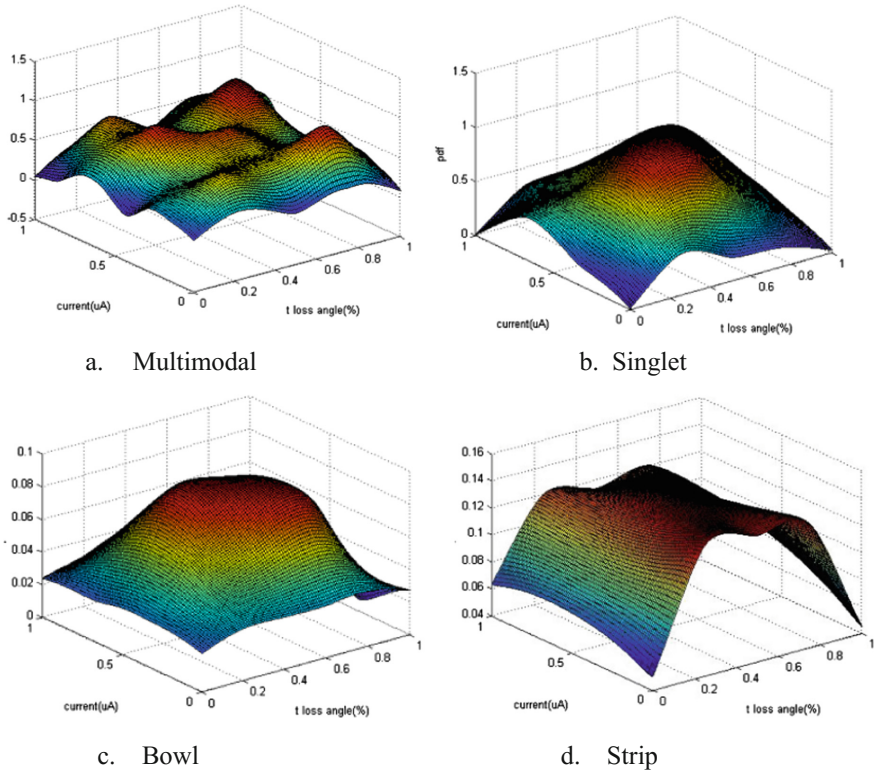
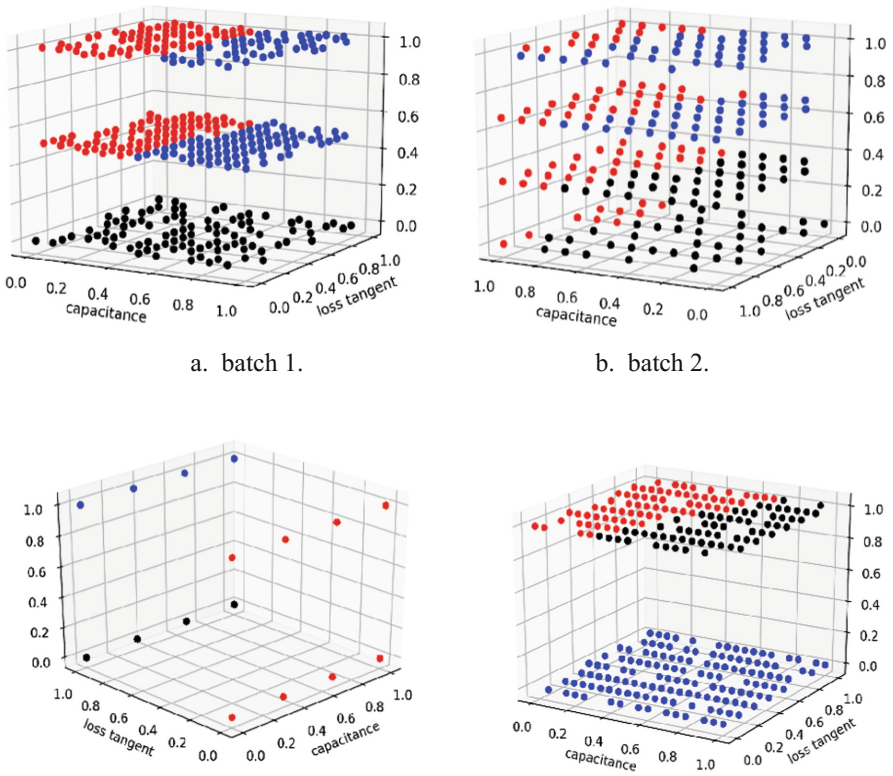


Fig. 2. Joint probability distribution of leakage current and minimum loss angle tangent

4.2 Data Clustering and Label Generation

According to the formed researches, only analyzing the relationships between variates are not enough to distinguish the illusory data, hence we adopt optimized k-means [13]

algorithm to dig out the inner features. Figure 3 is the clustering results on four batches datasets and these four batches all have six hundred records. The points in batch 1, 2 and 4 are very similar and they are equidistribution, but the points in batch 3 cluster in some fixed points, which illustrates the value of some attributes in batch 3 are same. The distribution in batch 3 is much similar to the crowd psychology phenomenon in social psychology. The crowd behavior is heavily influenced by the loss of responsibility of the individual and the impression of universality of behavior, both of which increase with the size of the crowd [2]. Hence we guess the third batch records are the illusory data. Then we use one third of real and illusory data as training datasets to optimize the kernel function and other datasets will be used to verify the accuracy of classifier.



a. batch 1. b. batch 2.

c. batch 3. d. batch 4.

Fig. 3. The clustering result of optimized k-means algorithm

4.3 Data Classification and Testing Behavior Analysis

Support vector machine [15] is one of the classical machine learning techniques that can still help solve big data classification problems, especially, it can help the multi-domain applications in a big data environment [12]. According to the clustering results, the real datasets are not linearly separable and the points will spread over some special points

uniformly. The parameters of kernel function will be changed based on the training datasets and then it will find a support vector to separate the points into different clusters. The Fig. 4 is the classification result on other two third of datasets, which is t for our clustering result and speculation. Using 2400 records to verify the accuracy of classification, 2115 points are same as their labels and the accuracy is about 0.88125, which means most records can get a proper label. Hence this classification can be used to distinguish illusory data from real data quickly if the features of datasets are similar to ours.

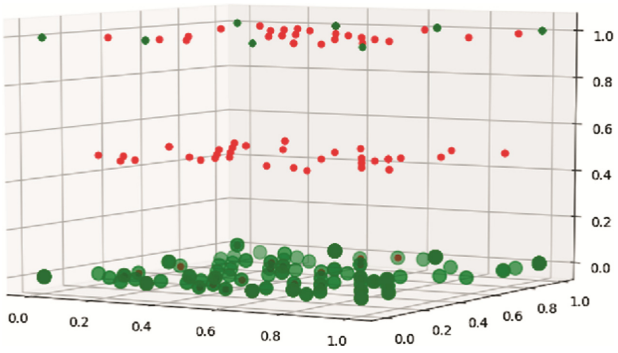


Fig. 4. The classification result of support vector machine.

From the data model, we can infer the testers' behavior. Most of the datasets are uniform distribution, only in batch 3, the points aggregate on some fixed points. This phenomenon is similar to the crowd psychology in social psychology. For example, if a tester is not willing to measure the variates of electronic components, they will see and copy others' records to decrease the risk of mistake. At first, they do not have an uniform criterion, so it will produce fluctuating results and be steady gradually. We still need more processes to confirm our conjecture.

5 Conclusion

In this paper, we propose a complete data mining framework to analyze the testing behavior and distinguish authenticity of testing records. This framework is generic and many datasets can use this framework to analyze their inner features and relevant behavior model. In order to increase the accuracy of clustering, we optimize the original k-means algorithm, which will make points spread uniformly. This classification can distinguish illusory datasets quickly, which will save much manpower and resource.

The proposed method still has some limitations. Firstly, the testing behaviors depend on different psychological activities, so every state of datasets are acceptable and we can only give a speculation which datasets may be illusory. Secondly the differences between real data and illusory are very tiny, and the common normalization method may decrease even eliminate these differences. In the future, we plan to implement more data mining algorithms in our framework to support other behavior analysis. And we can invite some professors to join the experiment to verify our clustering result. Meanwhile,

we will develop a data mining tool to detect the accuracy of testing datasets automatically, which will improve the quality of industry production.

Acknowledgement. This research is supported by the Shanghai Institute of Precision Measurement Project under Grand No. SAST2017-128 and the National Natural Science Foundation of China under Grant No. 61373030.

References

1. van der Aalst, W.M.P.: Process mining in the large: a tutorial. In: Zimányi, E. (ed.) eBISS 2013. LNBIP, vol. 172, pp. 33–76. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05461-2_2
2. Arain, S.M., Arain, A.M.: National Highways and Motorway Police in Pakistan: An Illuminative Study. Lulu.com, Morrisville (2016)
3. Bouguettaya, A., Yu, Q., Liu, X., et al.: Efficient agglomerative hierarchical clustering. Expert Syst. Appl. **42**(5), 2785–2797 (2015)
4. Dai, W., Ji, W.: A mapreduce implementation of C4. 5 decision tree algorithm. Int. J. Database Theory Appl. **7**(1), 49–60 (2014)
5. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Mach. Learn. **40**(2), 139–157 (2000)
6. García, S., Luengo, J., Herrera, F.: Data Preprocessing in Data Mining. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-10247-4
7. Hawkins, S.H., Korecki, J.N., Balagurunathan, Y., et al.: Predicting outcomes of nonsmall cell lung cancer using CT image features. IEEE Access **2**, 1418–1426 (2014)
8. Kesavaraj, G., Sukumaran, S.: A study on classification techniques in data mining. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–7. IEEE (2013)
9. Mundada, M., Gawali, B., Kayte, S.: Recognition and classification of speech and its related fluency disorders. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **5**, 6764–6767 (2014)
10. Rumbell, T., Denham, S.L., Wennekers, T.: A spiking self-organizing map combining stdp, oscillations, and continuous learning. IEEE Trans. Neural Netw. Learn. Syst. **25**(5), 894–907 (2014)
11. Satyanarayanan, K.S., Srikanth, B., Murugesan, M.: Tree dataset extraction using HAC based algorithm (2016)
12. Suthaharan, S.: Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. ISIS, vol. 36. Springer, Boston, MA (2016). https://doi.org/10.1007/978-1-4899-7641-3
13. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised clustering by seeding. In: Proceedings of 19th International Conference on Machine Learning (ICML-2002) (2002)
14. Horn, J.T.H., Krokstad, J.R., Amdahl, J.: Joint probability distribution of environmental conditions for design of offshore wind turbines. In: ASME 2017 36th International Conference on Ocean, Offshore and Arctic Engineering. American Society of Mechanical Engineers, p. V010T09A068 (2017)
15. Kisi, O., Parmar, K.S.: Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J. Hydrol. **534**, 104–112 (2016)
16. Nazeer, K.A.A., Sebastian, M.P.: Improving the accuracy and efficiency of the k-means clustering algorithm. In: Proceedings of the World Congress on Engineering, vol. 1, p. 1–3 (2009)