International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:3, 2017

Data Quality Enhancement with String Length Distribution

Qi Xiu, Hiromu Hota, Yohsuke Ishii, Takuya Oda

***Abstract*—**Recently, collectable manufacturing data are rapidly increasing. On the other hand, mega recall is getting serious as a social problem. Under such circumstances, there are increasing needs for preventing mega recalls by defect analysis such as root cause analysis and abnormal detection utilizing manufacturing data. However, the time to classify strings in manufacturing data by traditional method is too long to meet requirement of quick defect analysis. Therefore, we present String Length Distribution Classification method (SLDC) to correctly classify strings in a short time. This method learns character features, especially string length distribution from Product ID, Machine ID in BOM and asset list. By applying the proposal to strings in actual manufacturing data, we verified that the classification time of strings can be reduced by 80%. As a result, it can be estimated that the requirement of quick defect analysis can be fulfilled.

***Keywords*—**Data quality, feature selection, probability distribution, string classification, string length.

I. Introduction

WITH the advancement of IoT technology and sensor diversification [1], manufacturing data analysis is expected to solve social problems. On the other hand, mega recall is becoming a deepening problem resulting in tremendous economic loss and safety issues [2]. Therefore, defect analysis, such as root cause analysis and abnormal detection, is becoming increasingly essential to prevent mega recall. However, quality [3] of manufacturing data collected in factories are too low to be analyzed directly. Specifically, manufacturing events are generated by various machines in factory. As a result, data of various types are collected, such as machine ID, operator name, material ID and product ID. Normally, manufacturing events generated by different machines have different schemas, thus, they have different column names and data types at each column. However, these manufacturing events are forcibly integrated ignoring their original schemas in order to avoid constant redesign of database schema. The necessity of redesign comes from the frequent change of data format generated by constantly updating machines.

As the result of integration ignoring the original data schemas, various data types are mixed in each column of MES DB. In other words, the data in MES DB lacks structural consistency [4]-[6] and are of low data quality [3], [7]-[11]. Therefore, data classification and organization are previously required before defect analysis. Actually, in certain factory we investigated, defect analysis takes about 2 weeks due to the time-consuming data classification performed by manual. However, product providers, which are customers of factories normally require defect analysis result within 7 days. As a result, there arises a problem that defect analysis cannot be finished in requested time. Therefore, fast and high precision classifier is required.

The conventional data classification methods can be categorized in two kinds, probabilistic method and deterministic method [12]. When classifying strings in manufacturing data, the conventional probabilistic method has low precision using only two features, character frequency and character position although it creates rules in a short time. On the other hand, the conventional deterministic method takes a long time to create the rules for high classification precision by using three features: Character frequency, character position and string length. Thus, there is trade-off between rule creation time and classification precision in conventional methods when classifying strings in manufacturing data.

To overcome the trade-off, we propose new data classification method to correctly classify strings in manufacturing data in short time. With our proposal, the manufacturing data can be correctly classified in a short time. Thus, defect analysis can be finished in time requested by product providers.

The rest of this paper is organized as follows: Section II describes the background and challenge of this research. Then, the proposed data classification method, String Length Distribution Classification method (SLDC) is introduced in Section III. The evaluation of SLDC and its result is explained in Section IV. Section V concludes the research.

II. Background and Challenge

A. Background

As the recent trend in manufacturing vertical, many kinds of data are collected by various sensors. The revenue of IoT sensors in manufacturing vertical is increasing at compound average growth rate (CAGR) of 38% [1]. Thus, more and more sensors should be installed in factories in near future. As the result, it can be estimated that collectable data in manufacturing vertical will be rapidly increasing. On the other hand, mega recall, defined as the recall of defective products in which the effected number is more than threshold level per call, is becoming more serious.

According to [2], mega recalls have resulted in tremendous economic loss and safety issue recently. In North America, the

Q. Xiu, H. Hota, Y. Ishiiare and T. Oda are with the Computing Research Department, Center for Technology Innovation-Information and Telecommunications, Hitachi, Ltd., Research & Development Group, 292, Yoshida-cho, Totsuka-ku, Yokohama-shi, Kanagawa-ken, 244-0817 Japan.

Qi Xiu is with the Computing Research Department, Center for Technology Innovation-Information and Telecommunications, Hitachi, Ltd., Research & Development Group, 292, Yoshida-cho, Totsuka-ku, Yokohama-shi, Kanagawa-ken, 244-0817 Japan (e-mail: qi.wa.xiu@hitachi.com).

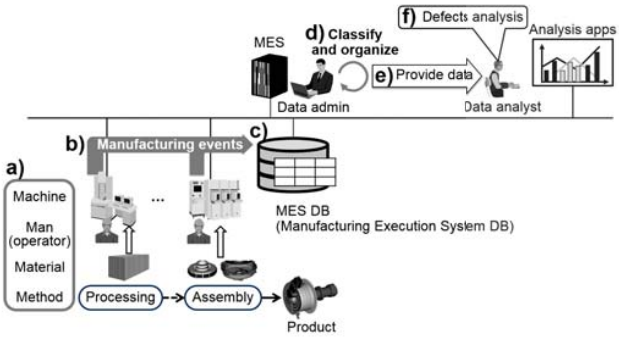

Fig. 1 Overview of a factory system

economic loss by mega recalls of automobiles is increasing at the CAGR of 24%. For example, by 2015, more than 34 million automobiles were recalled by defective airbag recalls, resulting in $2.7 billion loss. Furthermore, defective airbags have resulted in ten deaths and numerous serious injuries in the United States. Thus, mega recall is becoming a deepening social problem that must be solved. Therefore, defect analysis, such as root cause analysis and abnormal detection, is becoming increasingly essential to prevent mega recall. Under such circumstances, defect analysis utilizing manufacturing data is becoming more and more important. However, the data quality [3], [7]-[11] of manufacturing data collected in factories are too low to be analyzed directly. Fig. 1 shows the currently processing flow of manufacturing data in an actual factory we investigated.

a) In the factory, materials are manufactured by machines operated by operators according to manufacturing method.
b) In the above manufacturing process, manufacturing events are generated by various machines. Thus, data of various types are collected, such as machine ID, operator name, material ID and product ID
c) Manufacturing events are then stored in Manufacturing Execution System Database (MES DB). Although manufacturing events generated by various machines normally have different schema, they are forcibly integrated into MES DB ignoring original schemas.
d) Manufacturing events stored in MES DB are classified and organized by data admin.
e) The organized data are then provided to data analyst.
f) Finally, data analyst utilizes organized manufacturing data by applications such as abnormal detection system or root cause analysis system for defect analysis.

The reason manufacturing events stored in MES DB must be classified and organized previously is that there is data schema gap between table in MES DB and analyzable data table. Fig. 2 illustrates table in MES DB and analyzable data table as a) and b), respectively. As a result of forcibly manufacturing events integrating by ignoring their original schemas, columns names of a) table in MES DB are meaningless. Moreover, various data types are mixed in each column of a). Thus, a) table in MES DB lacks structural consistency [4]-[6]. On the other hand, b) analyzable data table must have original column names. Furthermore, b) analyzable table can only have one data type in each column.

To improve structural consistency and to bridge the gap, data classification and data organization are necessary to recognized the data types and reconstruct the original schema ignored in data integration. Normally, the classification of manufacturing events is performed by data admin using master data such as Bill of Materials (BOM) and asset list as hint to classify manufacturing events in MES DB. However, manual data classification results in long classification time while customer of factories, product providers normally require fast data classification in order to minimize the effect of recall. Specifically, customers of the factory we investigated requires defect analysis in 7 days after defective products occurring. So that the effect of recall can be minimized by stopping the shipment of automobiles effected by defective parts. However, as shown in Fig. 3, defect analysis takes about 2 weeks in the actual factory we investigated. In more details, data classification performed manually costs nine days itself, data classification performed by ETL tools costs half day and defect analysis costs another 6 days. Therefore, the data classification is the bottleneck of defect analysis flow and we must shorten data classification time to half day to meet the requirement of customer. Hence, it is necessary to provide fast classification method. On the other hand, high precision classification method even with incomplete master data are required because it is difficult to update master data matching the exact timing of frequent machine changing. Fig. 4 shows the case that master data such as asset list was lastly updated in October 2015 and new machine was added in February 2016. The new machine would be misclassified because it is not documented in master data and defect analysis would be failed. As the failure of defect analysis may result in serious economic loss for product providers or safety issue for users, we must provide high precision classifier with even incomplete master data. Therefore, the challenge of this research is to provide fast and high precision manufacturing events method with even incomplete master data.

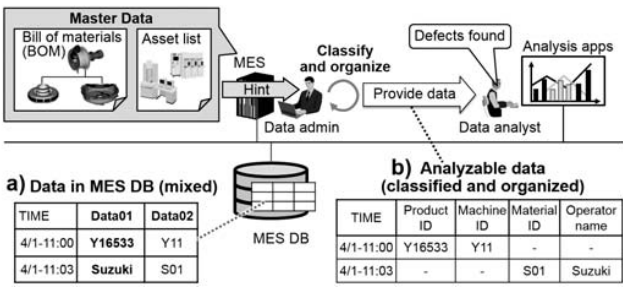

Fig. 2 Data schema gap between table in MES DB and analyzable data table

B. Conventional Data Classification Methods

In this subsection, conventional data classification methods are explained. However, only supervised classification methods automatically creating classification rules from master data are

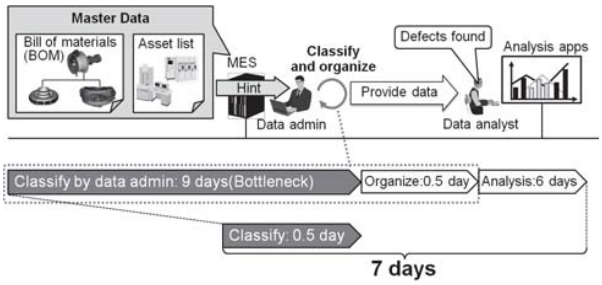

Fig. 3 Challenge of rule creation time

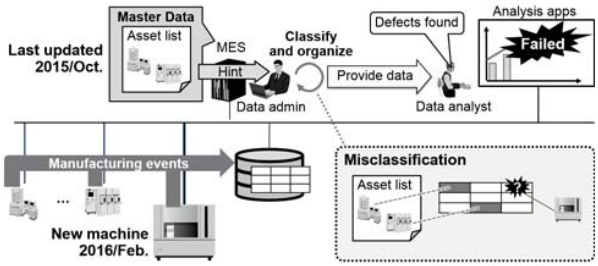

Fig. 4 Challenge of classification precision

discussed. Typically, there are 2 types of conventional data classification methods: Probabilistic data classification method and deterministic data classification method [12], which are explained as follows respectively.

1) Conventional Probabilistic Methods: Early in 1959, the method of estimating classification probabilities based on the statistical learning theory was introduced in [13]. Fig. 5 shows the conceptual model of the conventional probabilistic methods. As shown in the figure, the conventional probabilistic methods classify manufacturing events in three phases, feature extraction phase, rule creation phase and classification phase. In feature extraction phase, only 2 features are extracted from master data, character frequency and character position. In rule creation phase, based on character frequency and character position extracted from master data, the centroids of each group are calculated individually. In classification phase, the similarity between manufacturing event and centroid are calculated. Finally, manufacturing event are classified according to the similarity.

The conventional probabilistic methods create rules very fast because centroid is calculated by master data of only one data type. However, the precision is low by using incomprehensive featurez; only character frequency and character position.

Taking Naive Bayes Classifier [14] as an example, the actual similarity calculation process based on character frequency and character position is shown in Fig. 6.

a) When master data of Machine ID, 'Y11', 'Y22' are input. The probability distribution model showing the character appearance probability at each position is created based on master data in rule creation phase. At the first position, the character appearance probability of 'Y' is calculated as 100%. At the second position, character '1' and '2' have the appearance probability of 50%, respectively. Similarly, character '1' and '2' both have the probability of 50% at the third position.
b) In the classification phase, when manufacturing event 'Y21' is input, the similarity is calculated by multiplying the character appearance probability of 'Y' at the first position, '2' at the second position and '1' at the third position. The similarity of 'Y21' as Machine ID is therefore calculated as 25%.

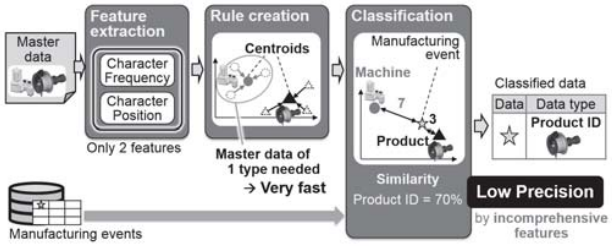

Fig. 5 Conventional probabilistic method

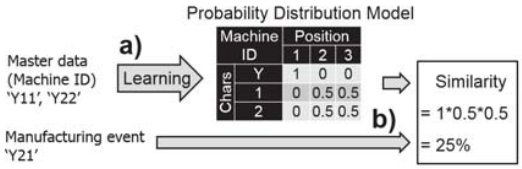

Fig. 6 Actual similarity calculation process of the conventional probabilistic method

For example, Tamr® 's data connection and enrichment platform [15] uses the conventional probabilistic method [16].

2) Conventional Deterministic Method: The conventional deterministic method automatically develops a set of classification rules and then apply these rules to classify target data. As example, [17]-[19] introduce the method of automatically generating regular expressions. Target data can be then classified by checking whether or not the they match previously created regular expressions.

As Fig. 7 shows, the conventional deterministic method classifies manufacturing events in three phases. Firstly, in feature extraction phase, in addition to character frequency, character position, string length are also extracted from master data as feature. By using comprehensive features, the precision of deterministic method is high. Then, in rule creation phase, boundary between different data types are defined. The rule creation time of deterministic method is long because complicated boundaries are calculated from master data of all data types. In classification phase, boundaries are used as rules to classify. In this case, manufacturing event is classified as Machine ID as it is in its boundary. For instance, IBM® InforSphere® QualityStage® [20] and Informatica® [21] use the conventional deterministic method.

C. Challenge

As Table I shows, there is trade-off between rule creation time and classification precision in conventional methods.

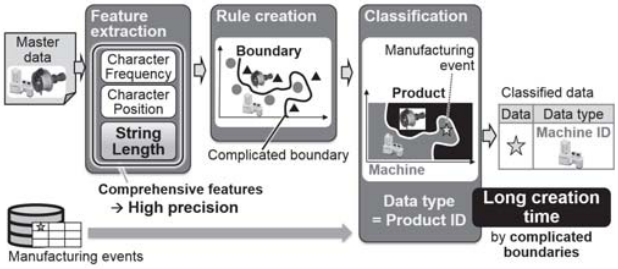

Fig. 7 Conventional deterministic method

The conventional deterministic method takes a long time to create the rules for high precision. On the other hand, the conventional probabilistic method creates rules in a short time but has low classification precision.

TABLE I
TRADE-OFF BETWEEN RULE CREATION TIME AND CLASSIFICATION PRECISION

	Conventional methods	
	Deterministic	Probabilistic
Rule creation time	☆☆★ Long	★★★ Short
Classification precision	★★★ High	☆☆★ Low

In more details, the conventional deterministic method has very long rule creation time. Because in rule creation phase, it needs master data of all types to define the complicated boundary between different data types. Hence it is not appropriate to be applied to manufacturing events with numerous data types. However, by using comprehensive features, the classification precision of conventional deterministic method is high. On the other hand, the conventional probabilistic method creates rules fast. Because in rule creation phase, it only needs master data of one data type to calculate centroid. However, by only using character frequency and character position as features, its classification precision is low.

When applying conventional probabilistic method to manufacturing events, such 2 data types of identifier, Product ID, 'Y10' and Machine ID, 'Y1', would be incorrectly classified as the same data type because there are 2 out of 3 common characters at the same position. Thus, the precision of the conventional probabilistic method is low by using incomprehensive features.

III. PROPOSAL OF DATA CLASSIFICATION METHOD: SLDC

In this section, we will explain our proposal, String Length Distribution Classification method, SLDC.

A. Key Points of Proposal

To correctly classify manufacturing events in a short time, we need to improve the precision of the conventional probabilistic method. For that purpose, we need to add comprehensive features used in the conventional deterministic method to probabilistic method.

We found that string length is a necessary feature when classifying manufacturing events. Because identifier is the key component to trace manufacturing events and string length is the indispensable feature classifying different types of identifiers.

Specifically, string length is indispensable feature classifying identifiers for two reasons. Firstly, identifiers of different data types have different string length. Because the string length of identifier depends on the total number of objects, such as product or machine needed to be identified. In mass production factory, the number of product is much larger than machines or materials. Therefore, Product ID need longer string length than Machine ID or Material ID.

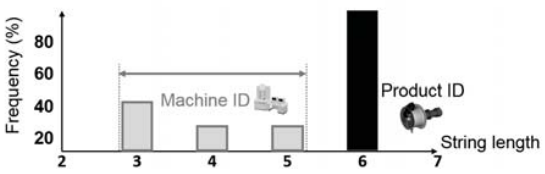

Fig. 8 String length variation in 1 manufacturing line

The other reason is that identifier of different types have different string length variation. As shown in Fig. 8, in one manufacturing line, Machine IDs have multiple length as they are sourced by multiple vendors while Product ID has only one string length as it is under unified product management.

In a word, string length is a dispensable feature to classify the key component of manufacturing events and therefore must be included as a feature in order to correctly classify manufacturing events.

B. Proposed Method

In this subsection, the flow of our proposal, SLDC will be explained. Since string length is the necessary feature when classifying manufacturing events, we add string length based learning to conventional probabilistic method as Fig. 9 shows.

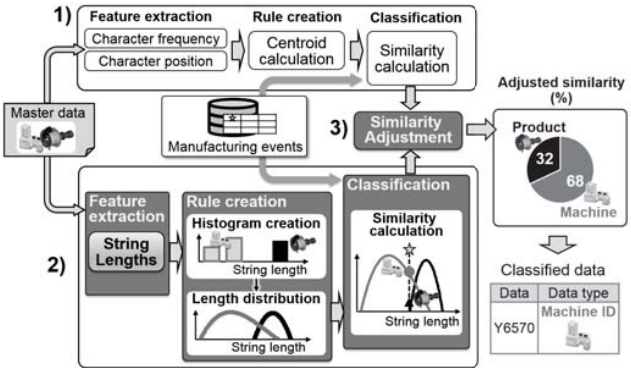

Fig. 9 Overview of SLDC processing flow

1) The same as probabilistic method, character frequency and character position are extracted from master data in feature extraction phase. Based on these features,

the centroids of each data type are calculated. Then, in classification phase, similarity is calculated based on the distance between the input manufacturing events and centroids of each data type.

2) In addition to that, SLDC extracts string length from master data as another feature. Then, in rule creation phase, the rule of string length is created. Length histogram is firstly generated based on the string length frequency of each data type in master data. Then length distribution is created to resolve string lengths not appeared in master data. In classification phase, similarity are calculated based on length distribution created.
3) Then, similarity calculated based on character frequency and character position is adjusted by similarity calculated based on string length. And manufacturing events are finally classified according to the adjusted similarity.

C. Example

Fig. 10 shows an example about how manufacturing event is classified by SLDC.

1) As the input, master data have machine id 'Y13', 'Y112' , product id 'Y65000', 'Y67000'. Meanwhile, manufacturing events 'Y6570', which is actually a machine id, is also input.
2) Centroids are calculated based on character frequency and character position of master data. Identifier has 'Y' at the 1st position and '1' at the 2nd position would be plotted near machine id. While identifier has 'Y' at the 1st, '6' at the 2nd and '0' at 4th to 6th position would be plotted near product id. As the input manufacturing event 'Y6570' has 'Y' at the 1st, '6' at the 2nd and '0' at the 5th position, it has much shorter distance with the centroid of product id.
3) On the other hand, length distribution is created based on length of master data. Length of Machine ID varies from 3 to 4 while length of product is ID fixed at 6. The input manufacturing events have the string length of 5, although length 5 has not appeared in master data. We can still calculate the similarity based on length distribution. Since product id is fixed at 6 in master data, the similarity of product id is only 5%. Machine id has 95% of similarity as its string length varies in master data.
4) Final process is the similarity adjustment by multiplying similarity calculated based on conventional method with similarity based on string length. After normalization, 'Y6570' is correctly classified to machine id as it has the similarity of 68%.

IV. Evaluation and Result

For the purpose of verifying the effectiveness of SLDC, we developed the prototype of SLDC and applied it to actual manufacturing events as evaluation. This section shows the details about the evaluation: The evaluation environment, evaluation result, benchmark with conventional data classification methods and discussion.

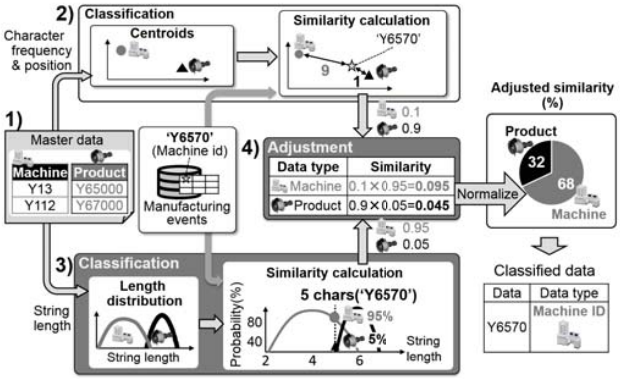

Fig. 10 Example of data classification by SLDC

A. Evaluation Conditions

In this subsection, the evaluation conditions are illustrated.

The experiment is executed on the virtual machine of OpenStack® . The hardware configuration of host is shown as:

- CPU: Dual Intel® Xeon® X5675 (3.07GHz/6cores) Hyper-threading was enabled
- Memory: 96GB (8GB (DDR3 1333 DIMM) * 12)
- Disk: 1TB * 2 (SATA® 6Gb/s/MLC) (Raid 0)
- Disk controller: LSI MR9261-8i
- Network: Intel® 82599ES 10GbE (for inter-VM network)
- Network interfaces: Intel® 82576 1GbE (for management)

The software of host is shown in as:

- IaaS: Openstack® Kilo
- Virtual Switch: Open vSwitch 2.3.2
- Virtualization API: Libvirt 1.2.12
- Hypervisor: QEMU® /KVM 2.2
- OS: Ubuntu® 14.04 (kernel 3.16.0)

The hardware configuration of the guest on which we performed experiment is shown as:

- CPU: 4VCPUs
- Memory: 8GB
- Disk: 80GB

The software of guest is shown as:

- IaaS: Openstack Kilo
- OS: Ubuntu 14.04 (kernel 3.13.0)

In the factory we investigated and collected target data, about 8000 products are manufactured everyday by 49 machines spreading on 4 production lines. More than 1M records of manufacturing events are collected every month, whose data size is about 250 MB.

The summary of target data is shown as:

- Classification target data: Actual manufacturing events
- Data size: 500 MB
- Number of columns: 213
- Number of records: 2.1 million

We used 2 months' manufacturing events for evaluation. The data size is about 500 MB and there are 213 columns and 2.1 million records.

Table II shows the samples of master data we generated to create rules. Each data type has 25 sample data. The sample of manufacturing events is shown as Table III.

TABLE II
EXAMPLE OF MASTER DATA

Product ID	Machine ID	Line no	Time
Y11541567	MA020110	VT-A-COV1	2014/12/15 1:15:00
S11541800	MA030110	VT-A-MAN1	2003/12/15 0:00:48
C11541026	MK010110	VT-M-SFT1	2003/12/15 0:46:00
V11531033	ML010110	VT-M-COV2	2015/11/11 4:16:00

TABLE III
EXAMPLE OF MANUFACTURING EVENTS FOR EVALUATION

Result no	...	DATA001	DATA002	DATA003	...
150401228	...	3276.7	LS110656	C11533190	...
150401230	...	257.4	-0.11	0.24	...
150401231	...	35.3	0.62	10000	...
150401233	...	3276.7	LS110656	C11533190	...
150401237	...	0	S11540129	V11540133	...

B. Comparison Target

As the comparison with SLDC, hybrid method, which is the combination of the conventional probabilistic method and the conventional deterministic method, is evaluated. The hybrid method is illustrated in Fig. 11. Firstly hybrid method uses the conventional probabilistic method to classify manufacturing events. As the precision of the conventional probabilistic method is only 63%, the misclassified data are then classified by conventional deterministic method. The rule creation time and classification time of both the conventional probabilistic method and the conventional deterministic method are evaluated.

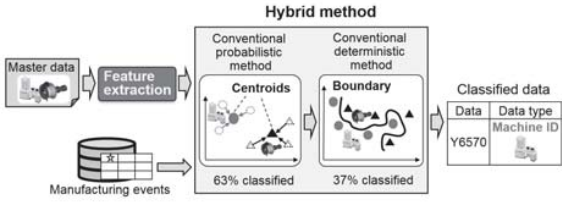

Fig. 11 Hybrid method

C. Evaluation Result

In this subsection, evaluation result will be explained.

As shown in Fig. 12, the horizontal axis shows the number of data types, and vertical axis shows the total time used for rule creation and classification. Hence, this graph shows the total time with the increase of data types. However, the data size for classification is fixed. There are 20 data types in the manufacturing events we used for evaluation. As shown in the figure, SLDC finished in 6 hours with the classification precision of 100%.

We also evaluated the total time of hybrid method as comparison. The hybrid method takes about 40 hours to classify manufacturing events, significantly exceeds half day's work time as required. Therefore, the total classification time can be reduced by 80% compared with conventional methods.

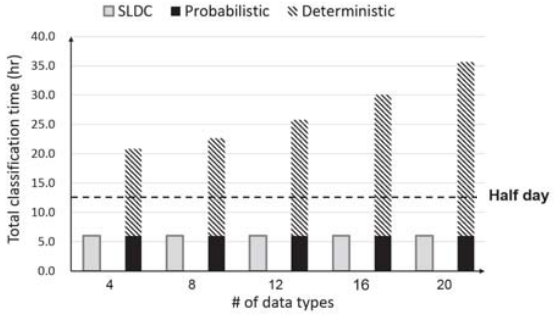

Fig. 12 Total classification time with the increase of data types

D. Benchmark

Classification method		SLDC	Conventional methods	
			Deterministic	Probabilistic
Features	Characteristic dimensions	Character frequency Character position String length	Character frequency Character position String length	Character frequency Character position
	Interpolation	Length distribution	-	-
Benchmark metrics	Time (total)	6 hr	45 hr	6 hr
	Rule creation	0.1 s	39 hr	0.03 s
	Classification	6 hr	6 hr	6 hr
	Precision	100%	100%	63%

Fig. 13 Comparison of SLDC with conventional methods

As shown in Fig. 13. SLDC is the only method which can 100% correctly classify actual manufacturing events in 6 hours. The first 2 rows show the features of each classification method. The first line shows features used in each classification method. SLDC and the deterministic method use character frequency, character position and string length three features, while the conventional probabilistic method only use character frequency and character position 2 features. The second line shows the interpolation used in each classification method, SLDC uses length distribution to resolve string length not appeared in master data while conventional methods have no method for interpolation. The next four lines show the classification time and classification precision of each method. As shown, SLDC can classify manufacturing events in 6 hours, much faster than 45 hours of the conventional deterministic method. On the other hand, SLDC has the classification precision of 100%, much higher than 63% of the conventional probabilistic method. Therefore, SLDC resolves the trade-off between rule/model creation time and classification precision in the conventional methods.

E. Discussion

1) The Change of Rule Creation Time with the Number of Data Types: As illustrated in Fig. 12, the total time of

SLDC and the conventional probabilistic method remains the same with the increase of data types while total time of the conventional deterministic method becomes longer with the increase of data types. The reason of this will be discussed as follows: The total time of deterministic method becomes longer with the increase of data types because the rule creation time increases quadratically with the number of data types. As mentioned before, in the rule creation phase, the conventional deterministic method defines complicated boundary using master data of all data types. Thus, the rule creation time of a single rule increase linearly with the number of data types as more master data are needed to define boundary differentiating different data types. As number of boundaries also increase linearly with the number of data types, the rule creation time of deterministic method increases quadratically. On the other hand, the total time of probabilistic method remains almost the same since the rule creation time increases linearly with the number of data types. In the rule creation phase, the conventional probabilistic method calculate the centroid of each data type. Hence, master data of one data type are needed in centroid calculation. Therefore, the creation time of single rule stays unchanged with the number of data types, and total rule creation time increases linearly. As the rule creation time remains within 1 secs even with 20 data types, the increase of rule creation time can be ignored and its total time is almost unchanged.

2) Using Proper Data Classification Method: As mentioned before, SLDC is designed to classify identifiers. Compared with conventional probabilistic method, the classification precision of SLDC was significantly improved by adding string length as another feature. However, when classification method is necessary for other purpose, such as entity resolution for data in customer management system, conventional probabilistic method should be used. The reason is discussed as follows.

For generic words, in the purpose of entity resolution [22], [23], 'Christ' and 'Christoph' are supposed to be classified as the same data type because they refer to the same person in the real world. On the other hand, 'Christ' and 'Chrome' are supposed to be classified as different data types since they are used to refer different entities in the real world. Using the features of conventional probabilistic classification method, character frequency and character position, 'Christ' and 'Christoph' are more likely to be classified as the same data type because they have more common characters at the same position than 'Christ' and 'Chrome'. The difference between their string length does not matter in this case. Furthermore, as shown in Fig. 14, when generic words and identifiers both exist, hybrid method of SLDC and the conventional probabilistic method can be used. Thus, SLDC is firstly used to extract identifiers and classify them. The rest data can then be classified by the conventional probabilistic method.

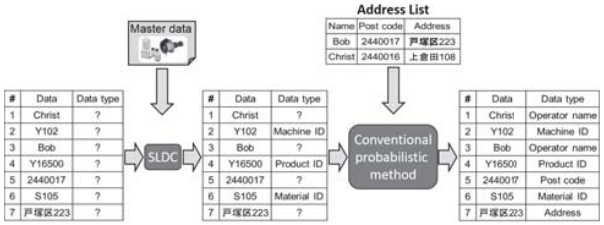

Fig. 14 Combination of SLDC and the conventional probabilistic method

V. Conclusion

In this research, we presented String Length Distribution Classification method (SLDC). The method resolves the trade-off between rule creation time and classification precision in the conventional data classification method when classifying strings, especially identifiers. We developed the prototype of SLDC and applied it to 2 months' manufacturing events collected from actual factory. The result shows that data can be correctly classified within 6 hours, reduced 80% compared with conventional methods. As a result, factories can reply defect analysis results in 7 days by utilizing SLDC to minimize the effect recalls.

As future work, we are applying SLDC to other products, other factories and other verticals to brush it up. In 2016, we plan to apply SLDC to at least three cases and 6 products in manufacturing vertical and financial vertical.

The second future work is implementing the methodology of automatic feature selection and include it into SLDC. In the example case, character frequency, character position and string length are the optimal feature subset to classify identifiers in manufacturing events. However, there is no assurance that these three features are also the optimal features for other factories and other cases. Hence, it is necessary to develop a automatic feature selection methodology. For example, scatter matrices [24][p280-p282] can be used as the criteria to evaluate the ability of feature for classification. The floating search methods [24][p286-288] can be used to optimize the subset of features.

References

[1] J. Rivera and R. V. D. Meulen. (2014, November 3). *Gartner Says the Processing, Sensing and Communications Semiconductor Device Portion of the IoT Is Set for Rapid Growth.* (Online). Available: http://www.gartner.com/newsroom/id/2895917 (accessed on 2016, October 31).

[2] National Highway Traffic Safety Administration (NHTSA), *Vehicle recall summary by year (1966-2014).* (Online). Available: http://www.safercar.gov/staticfiles/safercar/pdf/2014-annual-recalls-report .pdf (accessed on 2016, October 31).

[3] R. Y. Wang, and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," in JIMS 12, 4(1996), 5-34.

[4] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smisth, "A Framework for Information Quality Assessment," In JASIST, 58(12), 1720-1733.

[5] D.P. Ballou, H.L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," Management Science 31 (2), 1985, pp. 150-162.

[6] M. Jarke, Y. Vassiliou, "Data warehouse quality: a review of the DWQ project, Proceedings of the Conference onInformation Quality," Cambridge, MA, 1997, pp. 299-313.

[7] B. K. Kahn, D. M. Strong, and. R. Y.Wang, "Information quality benchmarks: Product and service performance," Communications of the ACM, 45, 4, 184-192, 2002.

[8] Y. W. Lee, D. M. Strong, B. K. Kahn and R. Y. Wang, "AIMQ: A methodology for information quality assessment," Information & Management, 40, 2 December, 133-146, 2002.

[9] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," Commun. ACM 45, 4, 2002.
[10] T. Margaritopoulos, M. Margaritopoulos, I. Mavridis and A. Manitsaris, "A Conceptual Framework for Metadata Quality Assessment," In DCMI 2008.
[11] M. Ge, and M. Helfert, "A review of information quality research – develop a research agenda," in Proceedings of the 12th ICIQ, Nov, 2007.
[12] Scott S., "Probabilistic Versus Deterministic Data Matching: Making an Accurate Decision," Information-management.com access in June 2009.
[13] H. B. Newcombe, J. M. Kennedy, S. Axford, and A. James. "Automatic linkage of vital records," in Science, 130(3381):954-959, 1959.
[14] A. K. Menon, O. Tamuz, S. Gulwani, B. Lampson, and A T. Kalai, "A machine learning framework for programming by example," in Proceedings of the 30th ICML, pages 187-95, 2013.
[15] "Tamr's data connection and enrichment platform data sheet," (Online) Available: http://www.Tamr.com/wp-content/uploads/2015/03/Technical_Data_Sheet_021915.pdf (accessed on 2016, October 31).
[16] M. Stonebraker, D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. Zdonik, A. Pagan, and S. Xu "Data curation at scale: The Data Tamer system," In CIDR, 2013.
[17] A. Bartoli, G. Davanzo, A. D. Lorenzo, M. Mauri, E. Medvet, and E. Sorio,"Automatic Generation of Regular Expressions from Examples with Genetic Programming," in GECCO, 2012.
[18] D. Lorenzo, E. Medvet, and A. Bartoli, "Automatic String Replace by Examples," in GECCO, 2013.
[19] A. Bartoli, G. Davanzo, A. D. Lorenzo, E. Medvet, and E. Sorio, "Automatic Synthesis of Regular Expressions from Examples," IEEE Computer, 2014.
[20] "IBM InfoSphere QualityStage data sheet," (Online). Available: http://public.dhe.ibm.com/software/data/sw-library/infosphere/datasheets/InfoSphereQualityStage.pdf (accessed on 2016, October 31).
[21] "Informatica Data Quality data sheet," (Online). Available: http://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/data-sheet/informatica-data-quality_data-sheet_6710.pdf (accessed on 2016, October 31).
[22] A. Doan, A. Halevy, Z. Ives, *Principles of data integration.* Waltham: Morgan Kaufmann. 2012, pp. 173-205.
[23] P. Christen, *Data matching concepts and techniques for record linkage, entity resolution, and duplicate detection.* Berlin-Heidelberg-New York: Springer, 2012, pp. 101-162.
[24] S. Theodoridis., K. Koutroumbas *Pattern recognition.* Burlington: Academic Press, 2008, pp. 261-322.