Data Pre-processing and Clustering Algorithm for Epidemic Disease Diagnosis Data

Yaoyao Sang
College of Information Science and Engineering
University of Jinan
Jinan Shandong China
email:1034168135@qq.com

Lianjiang Zhu
College of Information Science and Engineering
University of Jinan
Jinan Shandong China
email:ise_zhulj@ujn.edu.cn

Tao Du
College of Information Science and Engineering
University of Jinan
Jinan Shandong China
email:ise_dut@ujn.edu.cn

Shouning Qu
College of Information Science and Engineering
University of Jinan
Jinan Shandong China
email:qsn@ujn.edu.cn

***Abstract*—The paper mainly solves the problem of nonstandard tuberculosis data, and makes cluster analysis. It is an evaluation of existing work. The current epidemic disease data are huge and has great research value, but it is not easy to be directly used to discover knowledge. In this paper, the tuberculosis clinic data is pre-processed first; the ways of pre-processing mainly includes data cleaning, data integration, data transformation and data normalization. When processing the location information in the data, an innovative method of data weighting is proposed, which makes the complex medical data be numerical and normalized. Then, the novel unsupervised machine learning method Density peak clustering algorithm is used for clustering the data set and prove the validation of our method. This work can form clustering results and discover knowledge on this basis.**

Keywords—Tuberculosis clinic data; data normalization; location information; DPC.

I. INTRODUCTION

Nowadays, the demand for intelligent medical and health care is increasing day by day. The application of big data analysis and mining in the medical field includes many directions, such as individual and group medical planning, disease management, remote patient monitoring, etc. [1]. The dataset obtained from the patient visit records is the most objective and can reflect the real situation. The transmission pattern and trend of infectious diseases can be analyzed from it by data mining. For health care workers, big data analysis and application is conducive to the prevention and treatment of epidemics. therefore, it is of practical significance to the intelligent processing and data mining of medical data. However, most of the existing medical datasets are not standardized and incomplete, and the data types are diversified. Correspondingly, the processing results are easily disturbed and negatively affected by noise value, missing value, outliers and different types of data. For data mining, low-quality data will lead to low-quality mining results [2]. Therefore, how to conduct scientific and standardized processing of massive medical data has become a hot issue, which is also the basis and necessary condition for mining valuable medical information. In order to solve the above problems, this paper proposes a set of numerical processing system for epidemic disease treatment data, including data cleaning, data integration, maximum and minimum normalization [3], and unique heat coding and weighted numerical method. In addition, it innovatively proposes a method of setting initial values and weighting them to highlight the differences, and weight the data with strong relevance to the dimension where the sub-type data are located, so as to scientifically convert them into values in the range of 0-1.

The remaining of this paper is structured as follows. Section I mainly introduces the background and significance of the subject. In Section II, Literature related to the research content is listed. In Section III, the method and process of data preprocessing are introduced, and the innovative geographic location processing method is mainly elaborated. In Section IV, the main content is the application of density peak algorithm in standardized data set. Section V mainly summarizes the content of the paper and look forward the future work.

II. RELATED WORKS

In view of the fact that most data streams are mixed attribute data, researchers have also proposed some algorithms to directly process mixed attribute data. Zhang et al.[4] proposed a hybrid algorithm combining fuzzy clustering with particle swarm optimization (PSO) for incomplete data clustering, and missing attributes are represented as intervals, which are based on Clustream [5], They proposed the micro-clustering histogram representation for the classification property of mixed attributes, and used Poisson process to model the arrival time of samples. For the clustering problem of mixed attribute data, the existing algorithm adopted the following methods. One is to convert the type attribute to numeric attribute. In this way, the data sets can be transformed into a numeric attribute dataset, which can then be processed by clustering the numeric attribute data. Shih et al. [6] converted classification attributes into numerical attributes according to the similarity of classification attributes, and then adopted k-means algorithm for clustering. In literature [7], classification attributes were converted into numerical attributes by the transformation method of spherical coordinates, and k-means algorithm was used to cluster the results. In this way, the dataset was transformed from mixed type to sub-type, and then processed by clustering method of sub-type attribute data. Rodríguez-Jiménez et al. [8] proposed new methods to generate the lattice concept with positive and negative information to be used as a kind of map of attribute connections. David and Averbuch [9] converted numerical attributes into classification attributes through CH index, and used spectral clustering to process datasets. In order to realize clustering analysis of the typed data of set-valued attributes, Giannotti [10] proposed a Trk-means algorithm based on Jaccard distance. However, there is no further analysis on the convergence of the algorithm. In order to improve the clustering algorithm for typed data, Cao et al. [11] designed and implemented the sv-k-modes clustering algorithm; this paper proposed a comprehensive clustering analysis for the data of subtype of set-valued attribute. Wangchamhan et al. [12] used Gower distance measurement to measure mixed attribute data in order to overcome the limitation that the k-means algorithm was unable to effectively process the data of different types. Chen and He [13] proposed an adaptive peak density clustering algorithm for mixed attribute data. According to the analysis of data attribute relationship, the mixed attribute dataset was classified into three types, including digital dominant data, classified dominant data and balanced data, and the corresponding distance measurement was designed. Ding et al. [14] proposed a density peak clustering algorithm based on entropy to process mixed data with fuzzy neighborhood. A new similarity measure was proposed for the numerical type with uniform standard, the similarity of numerical part is regarded as the whole and calculate the similarity of sub-type parts separately. This shows that it is challenging to process mixed attribute data by clustering algorithm alone, and most scholars focus on algorithm innovation to process data. There are few innovative points in data pre-processing, in particular, there are few studies on the relatively large amount of medical data. Therefore, this paper aims at making innovations in data preprocessing and proposes a weighted numerical processing method for geographical location data.

I. METHOD AND REALIZATION OF MEDICAL DATA PRE-PROCESSING

In this section, we will present the Data Processing Methods and Processes.

A. Data Processing Methods and Processes

Currently, thousands of records about desensitized TB patients are given, which are respectively based on six municipal cities in a province. Statistical analysis was conducted in each municipal city based on patients, including basic personal information such as gender, age and home address, as well as disease information such as treatment time, patient source, diagnosis location and result. The overall information is relatively comprehensive and complete, which could effectively complete subsequent data pre-processing and data mining.

Figure 1 is an overall flow chart of treatment methods for epidemic disease treatment data. Considering the diverse types of treatment data and the large proportion of categorical values, appropriate data pre-processing methods are selected to convert the original data into a numerical type. This step lays a foundation for subsequent data mining.

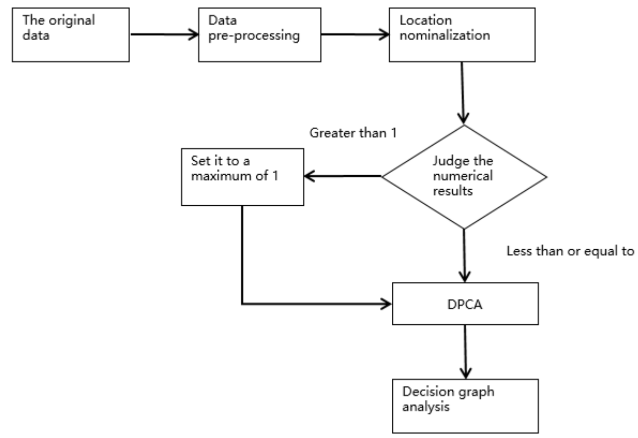

Figure. 1. Overall flow chart of treatment method for epidemic disease diagnosis data

B. Analysis of Data Pre-Processing Steps

The data pre-processing shown in Figure 1 is mainly divided into the following steps：

1) In the data pre-processing of the original medical data set, data cleaning was first carried out, which mainly to delete the irrelevant and duplicate data, and deal with missing values and outliers. According to the practical significance, if the missing value accounts for less than 2% and is not easy to fill, it can be directly deleted or ignored. However, some missing values will affect the machine processing results of the whole dataset. At this time, Lagrange interpolation method is selected for interpolation, and the formula of Lagrange interpolation is shown in (1).

$$L(x) = \sum_{i=0}^{n} y_i \prod_{j=0, j\neq i}^{n} \frac{x-x_j}{x_i-x_j} \tag{1}$$

Where, x_i is the x-coordinate of the known points, x_j is the function value corresponding to the known points, and the point x corresponding to the missing function value is substituted into the interpolation polynomial to obtain the approximate value L (x) of the missing value. Lagrange interpolation formula has a compact structure and is convenient in theoretical analysis. This method can model the missing data, which aims to be more objective and correct.

2) The original medical dataset is stored in the urban area, some attribute expressions are not matched in the same way, entity identification and attribute redundancy need to be considered. Entity recognition mainly deals with the data with different units. For example, the "first diagnostic area" in city A table and the "first diagnostic

 ISBN: 978-1-61208-846-4

unit" in city B table of the original dataset both refer to the medical institution where the patient was first diagnosed, which can be merged into the "first diagnostic unit". Attribute redundancy mainly refers to the repeated data caused by multiple occurrences of the same attribute or inconsistent naming. This kind of data is analyzed and detected before it is deleted. For example, the meaning of the attribute "date of birth" and "age" in a dataset are the same, and the "date of birth" attribute is deleted to reduce data redundancy.

3) In the original dataset, there are data similar to ethnic and occupational dimensions. In this paper, an equal-width discretization method is adopted for processing. Within the value range of data, about 5 discrete partition points are set, and the value range is divided into some discretized intervals. For gender (male, female), treatment classification (initial treatment, diagnosis and treatment), diagnosis result (positive, negative), and other obvious attributes of the classification, the method of 0-1 unique heat coding was used to nominalize them.

4) Age attribute of the original dataset up and the range is 0 to 93, the maximum minimum normalized for such data processing. According to the proportion, and scaling to a specific area to facilitate a comprehensive analysis, the purpose is to eliminate index between the dimension and the scope of influence, the use of the normalized formulas are shown in (2) below.

$$x^* = \frac{x - \min}{\max - \min} \tag{2}$$

Where, max is the maximum age and min is the minimum age, $(\max - \min)$ is range. Deviation standardization preserves the relationship existing in the original data and is the easiest way to eliminate the influence of dimension and data value range.

C. *Location Information Processing Method*

It seems reasonable to use latitude and longitude to represent geographical location, but it is not suitable in the overall data set. The reasons are as follows: first, latitude and longitude are coordinate values, which cannot meet the numerical requirements of 0-1 and the accuracy is not enough; second, the purpose of location numerical is not only to convert to 0-1, but also to be related to the reality, so other attributes with strong connection are used as parameters.

Geographic location information is classified data, which needs to be processed numerically to make it easy to handle. Therefore, this paper innovatively proposes a method to highlight the differences in the weighted processing of geographic location data. The main steps are as follows:

1) Divide the geographical location data into provinces, cities, regions and counties according to the administrative region from the largest to the smallest four levels, and different initial values are set respectively. The same grade is set to the same initial value, and the initial values in order of size are set according to the classification grade.

2) A new dimension is set for the attributes that are strongly correlated with the location data, and this attribute is used as the weighted factor z. In this paper, patients are taken as the unit and the number of patients in each hospital is used as the weighted factor.

3) Use the weighting formula to calculate the value of each location data, calculate the weighted radius on the basis of the initial value and make them all within the range of 0-1.

4) Compare and analyze the results of the initial location data and the weighted data by density map, and adjust the parameters reasonably according to the density analysis. The formula for calculating weighted value is shown in (3).

$$w_i = \frac{|\bar{z}-z_i|}{\bar{z}+z_i},\ R_i = \begin{cases} r + w_i, r = 0.5 \\ r - w_i, r = 1.0 \end{cases} \tag{3}$$

Among them, $\bar{z}$ as the average of number, z_i is the number of patients in the i i hospital, w_i represents the weighted value, the initial value r is graded according to the administrative region, provincial and municipal hospitals set to 0.5, district and county hospitals set as 1.0, Adjust the final numerical result R_i flexibly according to w_i, in the calculation, it will be greater than 1, we unified setting it to high of 1.0. Taking four different levels of diagnosis and treatment units as an example, the main parameters are shown in Table I

TABLE I. NUMERICAL TABLE OF DIFFERENT GEOGRAPHICAL LOCATION INFORMATION

Original geographic location	The initial rating value r	Number of visits z	Weighted value W	Numerical result R
People's hospital of A province	0.5	20	0.975848327	1
B city people's hospital	0.5	1239	0.138146912	0.638146912
County C center for disease control and prevention	1	1644	0.002377904	0.997622096
District D TB control station	1	2337	0.176381758	0.823618242

The initial value is set artificially in the light of the administrative region level, so as to reflect the difference in the weighted calculation. The key point is to adjust the weight value according to the actual situation. Figure 2 shows the histogram of the numerical results of the hospital in a certain city.

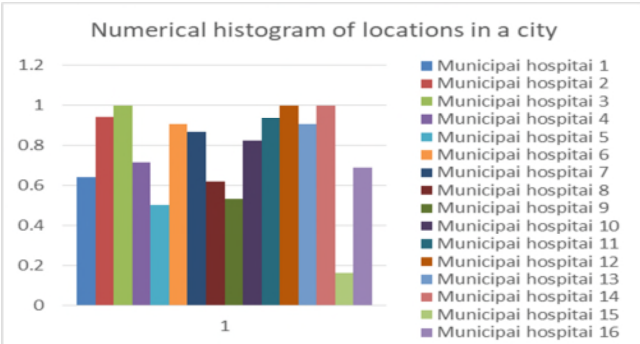

Figure. 2. Numerical histogram of locations in a city

We can see from Figure 2 that the processed position information data remains in the range of 0-1， and the standardization is realized.

II. REALIZATION OF DPC BASED ON MEDICAL DATA

In this section, we will present the Density Peak Clustering (DPC) Algorithm.

A. Density Peak Clustering Algorithm

DPC algorithm was proposed by Alex Rodriguez and Alessandro Laio [15] in 2014 and published in Science. The article "Clustering by fast search and find of density peaks" in Science mainly focuses on a kind of clustering method based on density. The main idea is to search for high-density regions separated by low-density regions. The DPC is based on the following assumptions: first, the density of the central point in a cluster is greater than that of the neighboring points; second, the distance between the center point of cluster and the higher density point is relatively large. Therefore, the DPC has two main quantities to calculate: first, the local density ρ; Second, the distance delta from the point of high density. Evidently, the center of the cluster with bigger ρ and delta relatively than others.

In order to verify that the pre-processing results of pulmonary tuberculosis medical data can be processed by machine learning, we used the popular unsupervised learning algorithm, DPC, to conduct experiments on MATLAB to extract valuable information from processed data.

B. Implementation of DPC

To reduce the time complexity and space complexity, in the original dataset, we selected the data of 11 dimensions with large influencing factors for verification: 1, the first diagnosis unit, 2, gender, 3, age 4, career, 5, treatment classification, 6, sources of patients, 7, the domicile belongs to, 8, patients diagnosis classification, 9, diagnosis, 10, reaction time, 11, diagnosis time. The decision graph is obtained by the density peak algorithm. Taking city A and city B as examples, the decision graph is shown in Figure 3 and Figure 4.

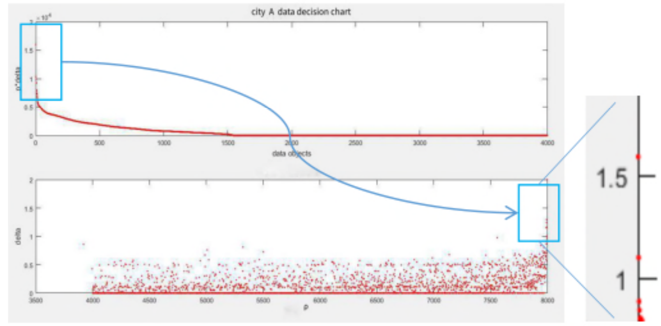

Figure. 3. Data decision chart of city A

Figure 3 is the decision diagram obtained by the density peak algorithm. The data points in the two boxes are corresponding, representing the center point in the cluster.

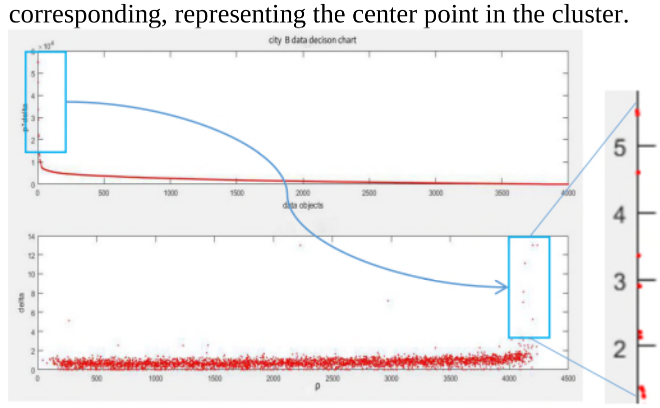

Figure. 4. Data decision chart of city B

Similar to Figure 3, cluster centers are data points with high density and high distance. It can be clearly distinguished in the decision diagram. According to the cluster center, the algorithm finds out the data in each cluster through continuous iteration and makes statistics. Figure 5 is the radar map of the first diagnosis unit after clustering in a certain city. It can be intuitively seen that among the four clusters, the two medical units with the largest number of patients are City A chronic disease prevention and treatment station, the other is B county tuberculosis prevention and treatment institute.

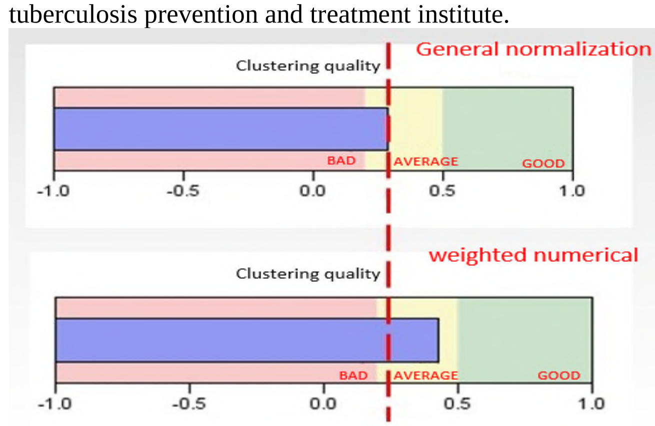

Figure. 5. Clustering quality comparison chart

By comparing the clustering quality of normalized and weighted numerical results of location attributes, it can be found that the clustering quality of innovative methods has obvious advantages from Figure 5, which also proves that the weighted numerical method is effective.

Figure 6 is the radar map of the first diagnosis unit after clustering in a certain city. It can be intuitively seen that among the four clusters, the two medical units with the largest number of patients are City A chronic disease prevention and treatment station, the other is B county tuberculosis prevention and treatment institute.

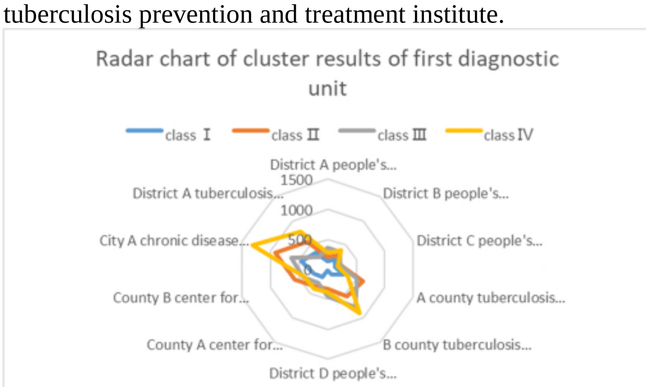

Figure. 6. Cluster radar map of the first diagnostic unit in a city

Figure 7 shows the occupational radar of patients and the reason for seeking medical treatment. It can be clearly seen that patients are mostly ordinary farmers, and they go to see a doctor after contact and infection It indicates that the pneumonia is widespread and most of them are farmers with low knowledge reserve and insufficient publicity, so that they may not be aware of it or do not know how to protect themselves from infection. In the future, knowledge discovery method will be used to analyze the statistical information of the decision graph, such as association rule algorithm, and combine the valuable attribute rules of the dataset to find the incidence rules of TB patients or treatment plans, so as to provide technical help for patients' treatment or doctors' decision making.

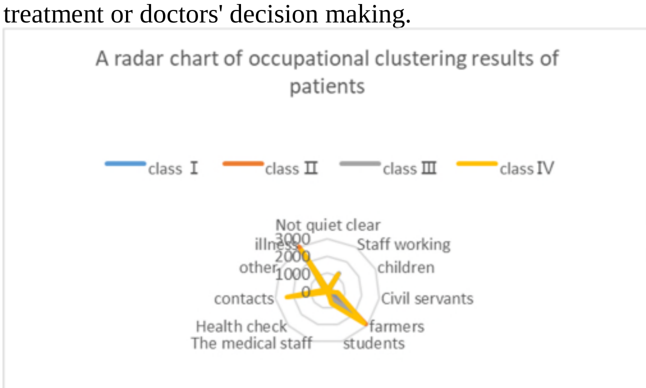

Figure. 7. Cluster radar chart of patients' occupation and reasons for seeking medical treatment in a certain city

III. CONCLUSION

In this paper, based on the treatment data of pulmonary tuberculosis patients in six urban areas, a data pre-processing method was proposed according to the characteristics of multiple attributes and non-standard. Moreover, a method to highlight the differences in the weighted processing of geographical location classification data was innovatively put forward, the original mixed datasets processing as a single numeric datasets. After that, the DPC algorithm was used to clustering data to obtain some valuable information, such as age or occupation distribution. Cluster centers and outliers were obtained by analyzing the decision graph. In future work, knowledge discovery method will be introduced, and association rules will be analyzed on the basis of clustering to discover tuberculosis treatment rules or other hidden knowledge.

ACKNOELEDGMENT

Thanks for the data support provided by Professor Li Huaichen and the respiratory team of Shandong Provincial Hospital, and thanks for the previous research foundation of Zhang Rui in the research group, which provided inspiration for the whole research and innovation.

REFERENCES

[1] Z. J Wang, S. W Mao, L. Y Yang, and P. P Tang. "A Survey of Multimedia Big Data." [J]. China Communications, vol. 15(01), pp. 155-176, 2018.

[2] M. D Li, H. Z Wang, and J. Z Li. "Mining conditional functional dependency rules on big data." Big Data Mining and Analytics vol. 3(01), pp. 68-84, 2019.

[3] G Xiang, and W Fang. "The research of Data Integration and Business Intelligent based on drilling big data." International Conference, pp. 64-68, 2017.

[4] L Zhang, Z Bing, and L Zhang. "A hybrid clustering algorithm based on missing attribute interval estimation for incomplete data." Pattern Analysis & Applications, vol. 39(1), pp.77-84, 2015.

[5] C. C Aggarwal, J Han, and R Ctr. "A framework for clustering evolving data streams." Proceedings 2003 VLDB Conference, pp.81-92, 2003.

[6] M Shih, J Jheng and L Lai. "A two-step method for clustering mixed categorical and numeric data." Tamkang Journal of Science and Engineering, vol. 13(1), pp.11-19, 2010.

[7] B. R Fatima, and J. L Diez. "Geometrical codification for clustering mixed categorical and numerical databases." [J]. Journal of Intelligent Information Systems, vol. 39(1), pp. 167-185, 2011.

[8] JM Rodríguez-Jiménez, P Cordero, M Enciso, and A Mora. "Data mining algorithms to compute mixed concepts with negative attributes: an application to breast cancer data analysis." [J]. Mathematical Methods in the Applied Sciences, pp. 4829-4845, 2016.

[9] G David, and Averbuch. "Spectral cat: categorical spectral clustering of numerical and nominal data Pattern Recognition." vol. 45(1), pp. 416-433. July 2011.

[10] Giannotti, C Gozzi, and G Manco. "Clustering Transactional Data." European Conference on Principles of Data Mining & Knowledge Discovery Springer, Berlin, Heidelberg, 2002.
[11] F Cao, J. Z Huang, and J Liang. "A fuzzy sv-k-modes algorithm for clustering categorical data with set-valued attributes." Applied Mathematics & Computation, vol. 295, pp. 1-15, Sep 2016.
[12] T Wangchamhan, S Chiewchanwattana, and K Sunat. "Efficient algorithms based on the k-means and chaotic league championship algorithm for numeric, categorical, and mixed-type data clustering." Expert Systems with Application, vol. 90(dec.30), pp. 146-167, 2017.
[13] J. Y Chen, and H. H He. "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data." Information Sciences. pp. 271-293, 2016.
[14] S. F Ding, M. J Du, T. F Sun, X Xu, and Y Xue. "An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood." Knowledge-Based Systems, pp.133, July 2017.
[15] A Rodriguez, and A Liao. "Clustering by fast search and find of density peaks." [J]. Science, vol. 344(6191), pp. 1492-1496, 2014.