The optimal values for learning set are the following: confidence threshold = 0.86, support threshold = 0.02, M = 2. With these values we achieved F1 score value 0.69 (precision = 0.89, recall = 0.57). Applying these values to test set we get F1 score value 0.68 (precision = 0.80, recall = 0.60). It indicated that learning model is neither under, nor overfitted.

The next step was to add random Medline articles in test set with pharmacology articles. We've selected 62 abstracts from *British Journal of Pharmacology*. In this case F1 score value still is 0.69 (precision = 0.81, recall = 0.60).

Conclusions and further works. The suggested method of DDI articles classification demonstrates its stability with different conditions, but its accuracy should be significantly improved. If the value of precision is high enough, the value of recall is low. In this case “low” and “high” estimates mean algorithm's applicability to practical tasks. 19% of incorrectly classified DDI articles can be filtered by human editor but 40% of articles with DDIs lost by classification algorithm are still too much. The following possibilities for accuracy improvements are considered:

- filtering of drugs names;
- using the number of drug names as additional parameter;
- increasing number of DDI articles in learning set;
- including DDI articles from other sources to learning set.

Acknowledgement. We would like to thank factsandcomparisons.com for granting us temporary access to their database.

REFERENCES

1. Facts & Comparisons. - 2014. - URL: http://www.factsandcomparisons.com/facts-comparisons-online/
2. Medscape from WebMD. Drug interaction checker. - 2014: URL: http://reference.medscape.com/drug-interactionchecker
3. U.S. National Library of Medicine. MEDLINE. - URL: http://www.ncbi.nlm.nih.gov/pubmed.
4. Stanford CoreNLP (software). - 2014. - URL: http://nlp.stanford.edu/software/corenlp.shtml.
5 Zhang Yun-tao. An improved TF-IDF approach for text classification // Journal of Zhejiang University Science. - 2005. - Springer.

BIOINFORMATIC SERVICE OF DECISION-MAKING BASED ON CLOUD COMPUTING

Gerget O.M., Mileshin A.A.
Tomsk (National Research Tomsk Polytechnic University)

Currently, biology and medicine is rapidly moving away from the verbal description and are based on mathematical models and information technology . Considerable difficulties of quantitative characteristics in the dynamics of their predetermined characteristics and properties, structural and functional complexity , variability parameters for the description of the biosystem , nonlinearity characteristics , incomplete and unclear description of the objects of study . Successful solution of health problems is impossible without the creation of appropriate information systems. One of the most difficult and time-consuming process of designing information systems is the identification of patterns of existing data sets . It does not always end well, because databases contain diverse and sonically contradictory and incomplete information .

Most of the currently available information technology oriented tend to solve specific practical problems and are compartmentalized , complex, expensive , making them unsuitable for

mass use in medical institutions. In this regard, the authors developed a system that allows you to identify patterns of temporary changes biosystem indicators based on statistical analysis and includes such important projects as the restoration of missing data , detecting the presence of seasonal rhythms , the allocation trends in time series , seasonal decomposition .

The purpose of research is to identify patterns of temporal change biosystem indicators based on statistical analysis .

Structure information of the medical system

To implement this goal in medical information system developed services using parallelized algorithms. Amongthem :

1. Service to identify trend-cycle component

To assess the availability of seasonal rhythms in the time series was used autocorrelation function and its graphical representation - correlogram . From the correlogram analysis can reveal the structure of the series. If the highest correlation coefficient appeared first order, then researched series contains only a tendency , if the autocorrelation coefficient of order h, then the series contains the cyclical fluctuations with periodicity h time points. The sequence of autocorrelation coefficients with offsets 1 , 2 , 3, etc. called autocorrelation function , whose values are in the range [-1 , 1] . Autocorrelation function should be used to highlight in the time series of the trend and seasonal components.

2. Service allocation trends in time series

Checking for non-random component is reduced to testing the hypothesis of the immutability of the average value of the time series with the runs test . When using it, you want the median y_{med} of time series , and the formation of " series " of the pros and cons of the following rule:

$$y_t = \begin{cases} +, y_t > y_{med} \\ -, y_t < y_{med} \end{cases}$$

Elements of time series , equal y_{med}, thus obtained does not take account . By " series " is asequence of consecutive consecutive pluses or minuses. The presence of non-random componentin the time series is determined from the condition :

$$\begin{cases} v(n) > \left[\frac{1}{2}(n+2-1{,}96\sqrt{n-1}\right], \\ K_{\max} < [3{,}3(\lg n+1)] \end{cases}$$

where v (n) - the total number of series , *Kmax* - length of the longest of the series, [] - the integer part of the number. To build a trend used method of moving averages and exponential smoothing method .

The method of moving averages is the following: first, determine the number of observations included in the smoothing interval . Then, the average value in the observation interval of smoothing by the formula:

$$\bar{y}_t = \frac{1}{m} \sum_{i=t-\frac{m-1}{2}}^{t+\frac{m-1}{2}} y_i ,$$

where *m* - number of observations included in the smoothing interval . Likewise is the smoothed value for other values , as long as the smoothing intervals will not last value of the time series .

An alternative approach to eliminate oscillations in a number of values is to use exponential smoothing method . Each smoothed value is calculated by combining the previous smoothed value and the current value of the time series. In this case, the current value of the time series taking into account the weighted smoothing constant , usually denoted α. Calculation itself is done by the following formula :

$$S_t = \propto y_t - (1 - \propto) S_{t-1}$$

where S_t- current smoothed value ;

y - the current value of the time series ;

S_t- 1 - the previous smoothed value ;

$\propto$ - smoothing constant.

The literature recommends a smoothing constant taking range from 0 to 1 , and in each case to select the most suitable value. [3]

3. *Service evaluation of seasonal decomposition*

To determine the seasonal component was developed algorithm seasonal decomposition data. Allocated on the basis of the trend of moving averages .Formed seasonal component – the difference or ratio between the original and the smoothed series. Calculated seasonal component - the average of all values of a number corresponding to a given point in the seasonal range.

The developed information system allows for processing of diagnostic data in parallel and implement a comprehensive approach to the diagnosis and prediction of the state of health of the human body , by combining into a single unit process analysis and control of information and the organization of operational data exchange in a single information space. Parallel data processing mode provides high utilization of computing resources by distributing a complex task into multiple computing nodes . Services include system design and allow authors to successfully diagnose diseases.

Findings

Checking for the seasonal component using correlogram showed that in some of the time-seriesdata present seasonal component .Seasonal Decomposition data series showed that they present a seasonal component . Analysis of seasonal indices showed that the changes of different indicators have laws that create a whole picture of the mutual changes of these parameters .

Trends isolated by two methods. It was concluded that the method of moving averages is more suitable for smoothing the time series, as it is more sensitive to changes in time series by the fact that when it is not recorded using the previous values of the smoothed row.

Conclusion

Time series analysis conducted in this paper allows us to represent the behavior of blood chemistry parameters in healthy people. Seasonal decomposition was carried out data series that gives an indication of changes in the patterns of these indicators in a certain period . Also were built correlogram . All this allows us to represent some standard behavior of these indicators over time to assess the state of the sick people and to evaluate the effectiveness of the treatment.