Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 3; April-June, 2016, pp. 169-173

http://www.krishisanskriti.org/Publication.html

Performance Evaluation of different Data Mining Techniques for different Medical Data Sets

Pardeep Kaur[1] and Anil Kumar[2]

[1,2]Computer Engineering and Technology Department, Guru Nanak Dev University Amritsar, Punjab
E-mail: [1]princi.thind@gmail.com, [2]anil.gndu@gmail.co

Abstract—*Precision prediction has long been a complicated trouble with the existence of a big number of missing information in the dataset. Several designs offer with this specific challenge have both removed the missing information from the data collection as called as event deletion and applied some various ways to load that missing data. This report targets a story combining Prediction Product with missing information assertion to analyze several techniques and use Easy K-means clustering and apply the most effective one to an information set. The main goal of this report would be to propose a fresh cross multi-class SVM based on missing price imputation. The further advancement will be done by using information parallelism approach of similar computing.*

Keywords*: Data mining, medical data sets, Data mining techniques in medical, Data mining applications*

1. INTRODUCTION

Data Mining (DM) is the procedure that obtaining new styles entrenched in large information sets. It employs these details to build predictive models. DM is found in data, databases, and artificial intelligence. The medical care market used information mining because they yield big and complicated amounts of data. DM can create data that may be useful to all stakeholders in medical care, including patients by pinpointing of use solutions and best practices. The main DM activities include explanation and apparition, looking for relations between information things, bunch information into units of same documents which method known an information classification, clustering, forecast based on traits that may be produced from information, etc.

Data Mining Process in the info growth knowledge method, the data mining strategies are used for finding styles from data. The styles that may be exposed depend upon the data mining projects apply. There are two types of data mining projects are detailed knowledge mining projects which explain the most popular homes of today's knowledge, and analytical knowledge mining projects that effort to accomplish predictions centered on presented data. In data mining, there are many medical care purposes and currently effectively established as analyzing treatment effectiveness, medical care management, the analysis of associations between people and providers of treatment, pharmacovigilance, scam and abuse detection. Despite the understandable benefits, there occur several constraints and difficulties in changing DM analysis techniques. DM can be confined by the accessibility to data that usually is spread in different options as scientific, administrative, insurers, laboratories, etc. Data may be unpredictable, corrupted, noisy or incomplete. Data mining of medical data involves specific medical knowledge in addition to familiarity with data mining technology.

1.1 Data mining applications

1. Data mining used in Healthcare – The accomplishment of healthcare knowledge mining handles on the accessibility of spotless healthcare data. It is essential for the healthcare market to appear into how knowledge can be greater caught, organized, stored and mined. In medical care, knowledge Mining is useful for the diagnosis and treatment of diseases and to spot the relationship that develops among many diseases. As healthcare knowledge aren't limited by just quantitative knowledge ,it can also be essential to examine the utilization of knowledge mining to grow the range of what medical care knowledge mining may do.

2. Data mining is used for marketing analysis-This approach is used in MBA(Market Basket Analysis). In market basket analysis when a customer wants to purchase some goods, then the data mining approach provide help to discover relations among various products which the buyer put within their buying cart or baskets. Here the discovery of such relations can be determined which grow the company techniques. The retailers use the data mining techniques to recognize the clients buying the pattern .In in this manner that approach is used for profits of the company and also helps to recognize the behavior of clients.

3. The data mining is used emerging trends in the education system –In the subject of knowledge, data mining is incredibly used and is a climbing field. As each year millions of students are enrolled across the country with a big number of larger knowledge aspirants, with knowledge mining technology will help linking understanding gap in larger educational systems. Data Mining assists in identifying not known habits, associations, and defects from educational knowledge and

may increase choice creating operations in larger educational systems. That improvement may give several advantages.

4. Usage of Data Mining in different areas of manufacturing engineering- When knowledge is gathered from production system it's used for various applications like to get the mistakes in the info or solution, to enhance the look strategy, to help make the high-quality solution. The brand new strategy was planned as CRISP-DM that will supply the higher level aspect steps of directions for utilizing the knowledge mining in the engineering.

5. In language research and language engineering: -Often a linguistic data will become necessary about a text. A linguistic profile that contains a lot of linguistic characteristics could be created from text record immediately applying information mining. That approach found very successful for authorship evidence and recognition. This technique verifies immediately the writing is of indigenous quality.

1.2 Medical data sets

Classification of knowledge from the College of Colorado, Irvine (UCI) machine understanding knowledge collection repository was executed to evaluate the effectiveness of the hybrid classifier on real-world knowledge, and to help contrast with other classifiers. The data models useful for that evaluation are described in detail in and at the UCI internet site .

1. Hepatitis- That knowledge includes 19 detailed and medical check effect values for 155 hepatitis individuals. The 2 lessons, children, and individuals for whom hepatitis proved final, are strongly unbalanced—123 samples fit in with the heir school while 32 fits in with the final class. The information includes qualitative, along with equally constant and distinct valued quantitative features. Several characteristics haven't any missing values while the others have as many as 67 missing values out of 155 samples. The little trial measurement and incompleteness with this knowledge set are typical of many medical classification problems.

2. Pima- Diabetes examination data for indigenous American girls of the Pima heritage, outdated 21 or over . That data consists diagnostic data for 768 girls; 268 of these people tried good for diabetes while 500 tried negative. Six of the nine functions are quantitative and constant, consisting of varied medical test results. The rest of the two functions, an era in years and amount of occasions pregnant, are quantitative and discrete. There are no missing function prices in the data. The completeness and reasonable dimensionality of the data collection ensure it is suited to testing the ability of a classifier and function extractor to steadfastly keep up or raise classification precision while reducing dimensionality when there are fewer functions to perform with.

3. Thyroid – The data consists of 21 clinical test effects for a couple of people tried for thyroid dysfunction 15 of these functions are binary-valued while the other 6 are continuous. The training data include 3772 instances from the entire year 1985, as the screening data include 3428 instances from the next year. The data are grouped into two courses, consisting of the people that were perhaps not identified as having particular kinds of hypothyroid disorder. The 2 courses are very unbalanced: working out data include 3487 bad diagnoses and 284 positive, as the screening data include 3177 bad samples and 250 positive.

4. Breast cancer – The dataset is obtained from the university of Medical Sciences of Universidad Nova Delaware Lisbon with 163 regulates and 95 cases, them all individual Portuguese Caucasians. Of the 95 cases, 50 of these have tumour detect after menopause in women above 61 years, while another 45 have tumour detection before menopause, women in 50 decades old. The tumour form is ductile carcinoma.

1.3 Data mining techniques in medical

There are some commonly Data Mining techniques which are used in medical field are as follows:

1. Classification learning:- The educational field there are a couple of categorized cases as instruction set and utilize it for instructing the algorithms. Among the qualified calculations, categorization of the test information is taken on the cornerstone for extracting patterns and rule from working set.

2. Numeric prediction:- This can be a change of categorization understanding with the exemption that in the position of predicting the different class and then the end result is a numeric value.

3. Association concept mining:- The association and designs among many characteristics which produced and create principles. The rules and designs are utilized predicting the categories of the test data.

4. Clustering: - Group of similar instances is put into a cluster in the clustering . The difficulties and negatives taking into consideration of sorting equipment learning to first recognize clusters and assign a brand new example to these clusters.

5. Time series evaluation - In that, the worthiness of an attribute examined over a time an average of at continually spaced time intervals. For instance, depending upon the situations of someone, values of particular characteristics may be acquired on a daily or hourly basis.

6. Predictive information modeling –It's a significant information mining job to find out potential information claims on the cornerstone of past and current values. Forecasts may possibly be manufactured on the cornerstone of regression, time series examination, or several other approaches.

7. Visualization methods – It is a helpful method of acquiring styles in a medical information set. Spread diagrams in a

Cartesian airplane of two fascinating medical attributes may be used to spot fascinating subsets of medical information sets. Like, for center people fascinating subsets are available regarding blood sugar levels

2. RELATED WORK

Fear, Elise C., et al.(2002) [1] has purposed a real foundation for breast tumour identification with oven imaging is the division in dielectric attributes of normal and malignant breast tissues. Image construction methods are created to improve tumour responses and minimize untimely- and behind time clutter. Effective recognition circular tumours are achieved with both planar and cylindrical programs, and related efficiency steps are obtained.

Xiong, Xiangchun, et al. (2005) [2] discussed a Knowledge mining and statistics evaluation could function as look for useful information in large sizes of data. The three solutions for identify chest cancer are mammography, FNA, and medical biopsy. The location accuracy of mammography is from 67% to 78%, the accuracy of FNA is irregular with various from 66% to 99%t the accuracy of a medical biopsy is nearly 100%.

IShouman, Mai, et al.(2012) [3] is planned the accessibility of large levels of medical knowledge results in the necessity for powerful knowledge evaluation resources to get rid of good use knowledge. This paper discovers breaks in the investigation of cardiovascular disease examination and proposes a type to methodically close these breaks to find out using knowledge mining practices to cardiovascular disease treatment knowledge can give a reliable efficiency as that accomplished in diagnosing center disease.

Han, Jianchao et al (2008) [4]presented an Information mining techniques that thoroughly applied in bioinformatics to analyze biomedical data. The info preprocessing, including feature identification and collection, outlier elimination, knowledge normalization and exact discretization, aesthetic knowledge analysis, concealed relationships finding, and a diabetes forecast product structure is done in this.

Barakat, Mohamed Nabil H (2010) [5] mentioned a using support vector devices for the analysis of diabetes. Effects on a life diabetes dataset reveal that intelligible support vector machine supplies a capable software for the forecast of diabetes, where a comprehensible rule collection have already been produced, with forecast sensitivity of 93%, the reliability of 94%, and specificity of 94%.

Saiti, Fatemeh, et al (2009) [6] has proposed the Thyroid gland provides thyroid hormones to help the regulation of body's metabolism. The first is Hypothyroidism that relates to the creation of inadequate thyroid hormone and other is hyperthyroidism associated with the creation of extortionate thyroid hormone. These techniques count generally on strong categorization solutions to deal with obsolete and irrelevant features.

Temurtas, Feyzullah (2009) [7] described thyroid hormones produced by the thyroid gland support instruction of your body's metabolism A relative thyroid illness analysis recognized by using multilayer, probabilistic, and learning vector quantization neural networks is considered.

Sartakhti, Javad Salimi et al (2012) [8] has planned an analysis of hepatitis illness, that is really frequent and main illness, which is done with a device learning technique that hybrid support vector device and imitation annealing are studied. The obtained classification reliability of the method is 96.25%.

Bascil, M. Serdar et al (2012) [9] presented a hepatitis illness analysis with a proper model of the hepatitis information which is essential categorization problem. The outcome of the study was in contrast to the outcome of the last reports reported emphasizing hepatitis illness analysis.

Çomak, Emre et al (2007) [10] has proposed a determination support program that classifies the Doppler signals of heart device into two lessons normal and abnormal to guide the cardiologist. LS-SVM and BP-ANN are employed to categorize the produced features.

Sartakhti,et al. (2012) [11], proposed a novel a device understanding technique that hybridizes support vector device and simulated annealing for examination of hepatitis disease.Intensively investigated support vector device because of its a couple of exclusive benefits is successfully established as a predicting technique in current years. The purchased classification reliability of our technique is 96.25%.

Khan, Aurangieb, and Kenneth Revett (2004) [12] describe how hard collection principle can be properly used as a tool for examining somewhat complicated choice tables like the Diabetes Database The use of a genetic algorithm based hard collection way of classification of diabetes and reached successful rate on the test pair of 83% studies.

3. PERFORMANCE ANALYSIS

This important section depicts the prices of numerous parameter mean utter error, root mean squared error, true positive charge, false positive error, precision, remember, f-measure, ROC place, effectively categorized situations, wrongly categorized situations.

Table 1

DATA SETS	MAE	RMSE	TP RATE	FP RATE	PRECISION	RECALL	F-MEAS	ROC AREA
Breast cancer	0.4289	0.4779	0.66	0.66	0.435	0.66	0.525	0.5
Pima diabetes	0.451	0.4681	0.682	0.682	0.465	0.682	0.553	0.5

Hungarian -14-heart-disease	0.1906	0.3119	0.6	0.6	0.36	0.6	0.45	0.5
Hepatitis	0.3222	0.3804	0.83	0.83	0.687	0.83	0.753	0.5
Ecoli	0.1858	0.3092	0.86	0.386	0.149	0.386	0.215	0.5
Hypothyroid	0.0734	0.1924	0.921	0.921	0.849	0.921	0.883	0.5
liver-disorders	0.486	0.4891	0.624	0.624	0.389	0.624	0.479	0.5
lung-cancer	0.4111	0.4465	0.727	0.727	0.529	0.727	0.612	0.5

TP (True Positive): It denotes files that predicted as true and they definitely really true.

FN (False Negative): It denotes files that predicted as false and they were definitely true.

FP (False Positive): It denotes the files that were predicted as true and they were definitely false.

TN (True Negative): It denotes the files that are predicted as false and they were definitely false.

Accuracy: Accuracy shows the ratio of correctly classified samples to the total samples. It indicates the predictions that were made correct.

$$\text{Accuracy} = \frac{TP}{TP + FP + TN + FN} \quad (1)$$

Precision: Precision are the part of retrieved instances that are correct. It is calculated accuracy.

$$\text{Precision} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (2)$$

Recall: Recall is the division of related instances that are recover it is the calculation of unity.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

F-Measure: it combines and balance recall and precision. It is the compromise between recall and precision. When both are high

$$\text{F measure} = 2 * \frac{\text{precision.recall}}{\text{precision} + \text{recall}} \quad (4)$$

F-measure is high. The high value of F-measure indicates better is the algorithm.

Table 1. shows the results of evaluations, various data sets were used for classification and results indicates that when we load dataset in WEKA machine learning tool, hypothyroid show the best result in root mean square error parameter

TABLE 2

DATA SETS	Correctly CI	Incorrectly CI
Breast cancer	65.9794%	34.0206%
Pima diabetes	68.1992%	31.8008%
Hungarian -14-heart-disease	60%	40%
Hepatitis	83.0189%	16.9811%
Ecoli	38.5965%	61.4035%
Hypothyroid	92.1217%	7.8783%
liver-disorders	62.3932%	37.6068%
lung-cancer	72.7273%	27.2727%

Table 2. shows the results of evaluations, various data sets were used for classification and results indicates that when we load dataset in WEKA machine learning tool, the minimum value of correctly CI parameter is 39% in ecolic and in incorrectly CI minimum value is 7% in hypothyroid.

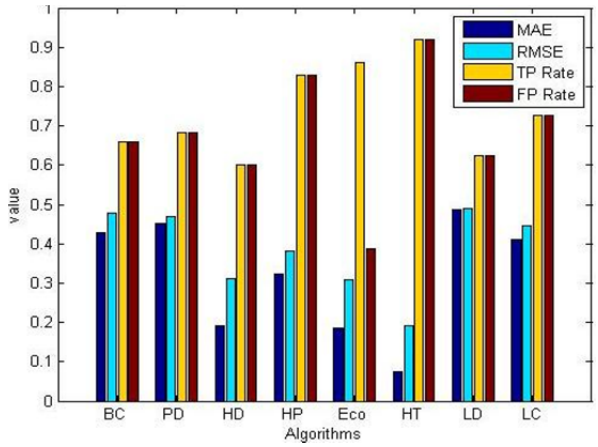

Fig. 1: Experimental Resuts 1

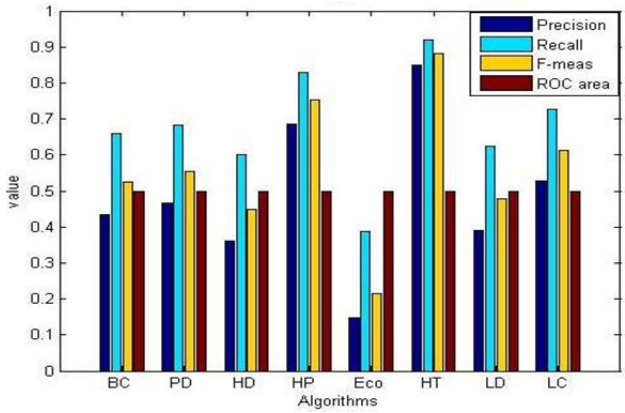

Fig. 2: Experimental Resuts 2

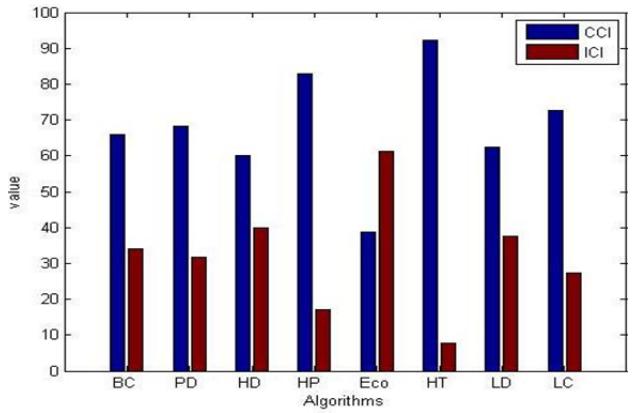

Fig. 3: Experimental Resuts 3

4. GAPS IN LITERATURE

1 The multiclass SVM classifiers style is really a further region for research.
2 while SVMs have good generalization effectiveness, they can be vastly measured in the test phase.
3 Moreover the features of SVMs from an efficient place of view they've some limitations. An essential practical question that's perhaps not fully determined is the choice of the kernel function parameter for Gaussian kernels the size parameter sigma and the worthiness of epsilon with epsilon insensible reduction task.
4 However, probably the most serious trouble with SVMs is the large algorithmic difficulty and extensive memory requirements of the mandatory quadratic development in large-scale tasks.

5. CONCLUSION

For locating attractive, perhaps useful, and previously not known habits from the enormous quantity of information is the method of information mining. They examined information that will be kept in an information storehouse. The major three techniques for information mining process are clustering, categorization and regression. A wide range of missing information contained in the information set with reliability has long been a hard problem. To address this dilemma, many designs might possibly erase the missing values from the information set as event deletion or they choose some other method to load that missing values. The near future area for study is the optimal style classifiers for multiclass SVM. The SVM good quality simplification performance, become really gradual in the testing stage. The kernel function variables variety is the disadvantage of SVM from a realistic level of view. The bigger difficulty of the algorithm is the key problem and quadratic programming in large-scale jobs needed more memory space. The overall goal of the report is to propose a fresh cross multi-class SVM based on missing price imputation.

REFERENCES

[1] Fear, E. C., Li, X., Hagness, S. C., & Stuchly, M. A. " Confocal microwave imaging for breast cancer detection: Localization of tumors in three dimensions". *Biomedical Engineering, IEEE Transactions on*, *49*,pp. 812-822,2002
[2] Xiong, X., Kim, Y., Baek, Y., Rhee, D. W., & Kim, S. H. "Analysis of breast cancer using data mining & statistical techniques". In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 200. Sixth International Conference on* IEE pp. 82-87,2005
[3] Shouman, Mai, Tim Turner, and Rob Stocker. "Using data mining techniques in heart disease diagnosisandtreatment." *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on*. IEEE, 2012.
[4] Han, Jianchao, Juan Carlos Rodriguez, and Mohsen Beheshti. "Diabetes data analysis and prediction model discovery using rapid miner." *Future Generation Communication and Networking, 2008. FGCN'08. Second International Conference on*. Vol. 3. IEEE, 2008.
[5] Barakat, Mohamed Nabil H., and Andrew P. Bradley. "Intelligible support vector machines for diagnosis of diabetes mellitus." *Information Technology in Biomedicine, IEEE Transactions on* 14.4 pp. 1114-1120,2010
[6] Saiti, F., Naini, A. A., Shoorehdeli, M. A., & Teshnehlab, M. " Thyroid disease diagnosis based on genetic algorithms using PNN and SVM." In *Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on* IEEE pp. 1-4, 2009
[7] Temurtas, Feyzullah. "A comparative study on thyroid disease diagnosis using neural networks." *Expert Systems with Applications* 36.1 pp.944-949,2009
[8] Sartakhti, Javad Salimi, Mohammad Hossein Zangooei, and Kourosh Mozafari. "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)." *Computer methods and programs in biomedicine* 108.2 pp. 570-579,2012
[9] Bascil, M. Serdar, and Halit Oztekin. "A study on hepatitis disease diagnosis using the probabilistic neural network." *Journal of medical systems* 36.3 pp.1603-1606,2012
[10]Çomak, Emre, Ahmet Arslan, and İbrahim Türkoğlu. "A decision support system based on support vector machines for diagnosis of the heart valve diseases." *Computers in Biology and Medicine* 37.1 pp. 21-27,2007
[11] Sartakhti, Javad Salimi, Mohammad Hossein Zangooei, and Kourosh Mozafari. "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)." 108.2 pp.570-579,2012
[12] Khan, Aurangieb, and Kenneth Revett. "Data mining the PIMA dataset using rough set theory with a special emphasis on rule reduction." *Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International*. IEEE, 2004
[13] Xing, Y., Wang, J., Zhao, Z., & Gao, Y. "Combination data mining methods with new medical data to predicting outcome of coronary heart disease." In *Convergence Information Technology, 2007. International Conference on* .IEEE pp. 868-872, 2007
[14] Kuo, W. J., Chang, R. F., Chen, D. R., & Lee, C. C. " Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images."*Breast cancer research and treatment*, *66*(1), pp.51-57, 2001
[15] Duch, Włodzisław, and Rudy Setiono. "Computational intelligence methods for rule-based data understanding." *Proceedings of the IEEE* 92.5 pp.77