IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Particle Swarm Optimization Based Support Vector Machine For Diabetes Mining

Ramandeep Kaur[1*], Prabhdeep Singh[2]

[1]Dept. of Computer Science and Engineering, Khalsa College of Engineering and Technology, Amritsar, India
[2]Dept. of Computer Science and Engineering, Khalsa College of Engineering and Technology, Amritsar, India

[*]*Corresponding Author: rgill4799@gmnail.com*

Available online at: www.ijcseonline.org

Accepted: 14/Aug/2018, Published: 31/Aug/2018

***Abstract*—** Data mining is the computational procedure for discovering routines within big files portions ("big files") pertaining to techniques in the intersection involving synthetic contemplating capability, unit learning, data, as well as collection programs. In this paper, we have proposed a new method in order to improve the accuracy of diabetes classification rate. The proposed technique have integrated Particle swarm optimization (PSO) with support vector machine (SVM) based machine learning technique. The proposed technique also verified by using the various standard diabetes classification data sets. The comparison drawn among the proposed and the existing technique based upon the various standard quality metrics of the data mining. Experimental results indicate that the proposed algorithm is more efficient than existing techniques.

***Keywords*—**Data Mining, Particle Swarm Optimization, Suppport Vector Machine, Diabetes Mining

I. INTRODUCTION

Data mining is the computational procedure for discovering routines within big files portions ("big files") pertaining to techniques in the intersection involving synthetic contemplating capability, unit learning, data, as well as collection programs. The whole goal involving the details exploration method is usually to extract data through the details fixed along with convert that right into a beneficial simple to comprehend shape for extra use. Form natural evaluation action, that includes files Loan Company and information administration factors, files pre-processing, item along with inference facts to consider, interestingness measurements, intricacy considerations, post-processing involving uncovered households, visualization, as well as online modernizing.

The true information pursuit job is usually the automated and also semi-automatic investigation involving major sums of info in an effort to acquire in the past mysterious, fascinating styles for instance teams of information (cluster analysis), abnormal documents (anomaly detection), as well as dependencies (association notion mining). This specific commonly consists of using databases procedures for instance spatial indices.

These kind of styles can be observed when a form of summary of your insight information and facts, and also can be employed throughout even more investigation and also, to get good example, throughout appliance discovering as well as predictive analytics.

To give an example, the info mining phase could discover various teams through your data, which can be used to obtain more genuine prediction benefits utilizing a decision, help system. Or your current information selection, information and facts prep, neither outcome meaning as well as reporting will be the principle information pursuit phase, but accomplish remain in the whole KDD method when even more steps.

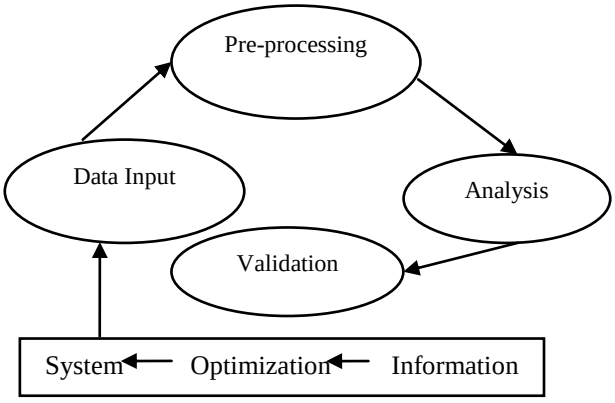

Figure 1. *Data* Mining *Process*

Details exploration is definitely the time period often utilized in pc science. Oahu is the approach used to acquire the actual

helpful details via big data fixed making use of various techniques. A variety of data exploration algorithms are utilized to acquire details via the data fixed including Group, Clustering, Aggregation and many more. The entire aim involving the data exploration approach is mostly to be able to acquire details via the data fixed as well as change the item into a superior easy to Understand variety which they can use further. Merely, getting helpful details via the data is actually called as data mining.

Data mining is principally used currently simply by firms using a substantial shopper concentration - retail store, economical, communication, and also advertising organizations. This helps they then to figure out connections between "central" aspects like selling price, merchandise placing, or maybe staff members ability, and also "exterior" aspects like economic indications, competition, and also shopper demographics. And, the idea helps these folks to discover have an effect on product sales, service delivery, and also corporate profits. Eventually, the idea helps these phones "soccer drills speed decrease" in to summary info to view aspect transactional data. With information exploration, a store could use point-of-sale information regarding shopper expenses to deliver specific marketing promotions based on a person's purchase history.

A. Abbreviations and Acronyms Data Mining In Healthcare

Medical industry currently builds massive amounts of complex info regarding clients, medical centers resources, ailment a diagnosis, electric client files, healthcare devices etc. These large amounts of information are an essential learning resource being processed as well as analyzed with regard to knowledge removal allowing aid with regard to cost-savings as well as making decisions. Data mining gives some sort of tools and methods that may be relevant to this particular ready-made info to find out secret patterns that include health pros an extra source of knowledge to create decisions. The actual decisions relax using medical professionals.

As a last paragraph of the introduction should provide organization of the paper, Section II contains the introduction of Particle Swarm optimization, Section III contain the related work, section IV explain the methodology with flow chart, Section V describes results and discussion, Section VIII concludes research work with future directions.

II. PARTICLE SWARM OPTIMIZATION

The Particle swarm optimization (PSO) is a Fuzzy C-Mean clustering formula is usually ways to show exactly how facts might be labeled in addition to grouped with group or even in any program [13]. This began by Dunn [14].In this kind of papers, utilizing Fuzzy c-means clustering formula background front things usually are segmented from your picture or even frames. This kind of formula mostly helps to portion your p whether it belongs to track record or even foreground. The sheer numbers of groups is generated good range of things from the frames. Using this wooly d indicates clustering formula centroid will probably be selected. First your centroid is usually picked out randomly good suggest in the pixels. The right centroid will probably be worked out immediately after obtaining the quality of pixel utilizing much iteration. Within this papers wooly c-means clustering strategy is employed for choosing your centroid good p along with the found sides utilizing the fresh border detectors formula [11].The pursuing formula displays how a wooly c-mean clustering procedure could be used to portion your front thing from your offered image/frame.

III. RELATED WORK

Cios et al. [1] deals with the particular exclusive top features of information mining by using healthcare data. He's discussed a variety of honest plus legalised areas linked to healthcare information mining just like information management, fear of legal cases, forecasted benefits, plus exclusive supervision issues. With this paper your dog stated which the numerical knowledge of approximation plus theory development in healthcare data is mainly not the same as all other information variety routines. Mitra et al. [2] comes with a market research on the obtainable materials on information mining utilizing gentle processing. A categorization offers been recently supplied predicated on various gentle computing devices and hybridizations put on, the data mining operate put on, plus the desire criterion identified by way of the unit. Inherited algorithms provide successful look for algorithms to pick out one, via mixed advertising information, predicated on a number of choice criterion objective function. Hard items are usually suitable to handle various types of concern with data. Bellazzi et al.[3] provides discussed which the prevalent option of brand new computational techniques plus resources pertaining to information exam plus predictive modelling necessitates healthcare informatics gurus plus experts to help systematically select the most appropriate approach to cope with controlled forecast problems. An enormous assortment of these methods needs common plus essential pointers that can help experts within the appropriate variety of knowledge mining resources, progression plus agreement with predictive types, combined with distribution with predictive patterns within just clinical environment. Palaniappan et al.[4] researched thatthe health care industry collects huge amounts of medical care information. These studies operate has created one Smart Soul Disease Prediction System (IHDPS) implementing information mining strategies, Choice Timber, Unaware Bayes plus Sensation problems Network. Success express that many approach has its own exceptional muscle with spotting the particular ambitions on the acknowledged mining goals. Making use of healthcare single profiles just like grow older, gender, blood pressure level plus sugar levels perhaps it will predict the particular likelihood of

people finding a coronary heart disease. Marungo, Fumbeya, et al.[5] introduced a information exploration program made to support the particular quick development of data-derived NTCP models. Prestashop exploits the normal healthcare workflow and information encoded by using a normal ontology. Mcdougal stated which the system referred to is a helpful information on the advance with irradiation oncology information exploration types especially plus local-level LHS components with general. Gholap, et al.[6] offers consist of your collaborative files mining procedure to offer multi-level evaluate coming from health test out data. The aim would be to examine success by simply collaboratively implementing various files mining techniques such as classification, clustering, along with connection procedure mining. General, the method seeks from getting information coming from health test out files to be able to increase model of health checks by simply creating peace of mind inside the final results implementing multi-level evaluate Nie, Liqiang, et al.[7] offers your story technique so that you can code a health data files by simply collectively employing local mining along with worldwide studying approaches, that are tightly hooked up along with mutually reinforced. Neighborhood mining efforts so that you can procedure the individual medical record by simply independently receiving the health strategies from the medical record itself and then maps these phones authenticated terminologies. Duan, K. B.et al.[8] offers consist of the latest feature choice procedure which relies on a backwards elimination treatment a lot like which put in place in assist vector equipment recursive feature elimination (SVM-RFE). As opposed to a SVM-RFE procedure, at each step, a consist of approach computes a feature ranking score coming from a mathematical analysis connected with bodyweight vectors connected with various straight line SVMs skilled with subsamples connected with the very first coaching data. Brameir.M et al. [9] Offers talked about a pair of methods of speeding connected with anatomical encoding approach. First an example may be the usage of a proficient algorithm which eliminates code. Subsequent one particular is a demotic approach to just about parallelize the system one processor.GP efficiency with health classification troubles can be as opposed coming from a benchmark data source having success received by simply nerve organs networks. Success demonstrates that GP functions equally in classification along with generalization. Prather J.C et al.[10] offers used the approaches of data mining (also called Understanding Uncovering in databases) to find interactions inside of a large professional medical database. They will illustrate a methods linked to mining your professional medical data source including files warehousing, files query& clean-up along with files analysis.

IV. METHODOLOGY

This section discusses the proposed algorithm.

Steps	Description
Algorithm parameters	A : Population of agents p_i : Position of agent a_i in the solution space f : Objective function v_i : Velocity of agent's a_i $V(a_i)$: Neighborhood of agent a_i (fixed) The neighborhood concept in PSO is not the same as the one used in other meta-heuristics search, since in PSO each particle's neighborhood never changes (is fixed).
Required attributes with values	Number of particles usually between 10 and 50 C_1 is the importance of personal best value C_2 is the importance of neighborhood best value Usually $C_1 + C_2 = 4$ (empirically chosen value) If velocity is too low → algorithm too slow If velocity is too high → algorithm too unstable
Particle update rule	$p = p + v$ with $v = v + c1 * rand * (pBest - p) + c2 * rand * (gBest - p)$ where*p*: particle's position, *v*: path direction, c_1: weight of local information, c_2: weight of global information, *pBest*: best position of the particle, *gBest*: best position of the swarm and *rand*: random variable
PSO Algorithm	$[x*] = PSO()$ $P = Particle\_Initialization();$ $For\ i = 1\ to\ it\_max$ $For\ each\ particle\ p\ in\ P\ do$ $fp = f(p);$ $If\ fp\ is\ better\ than\ f(pBest)$ $pBest = p;$ end end $gBest = best\ p\ in\ P;$ $For\ each\ particle\ p\ in\ P\ do$ $v = v + c1 * rand * (pBest - p) + c2 * rand * (gBest - p);$ $p = p + v;$ end end
PSD based K-means algorithm for training	1. Create a 'population' of attributes (particles) from developed clusters of data set X using K-means algorithm. 2. Evaluate each particle's position according to the objective function i.e. minimum root mean squared error (RMSE). 3. If a particle's current position is better than its previous best position, update it. 4. Determine the best particle (according to the particle's previous best positions). 5. Return optimized trained model

V. RESULTS AND DISCUSSION

For research and implementation the proposed technique is appraised using WEKA tool. The evaluation of proposed method is done on the origin of following parameters such as Accuracy, Mean square error and Sensitivity. The subsequent data demonstrates the comparison regarding response to diverse parameters. The result demonstrates the proposed solution provides improvement over active approaches. After the results, we compared the proposed solution against the current procedures.

A. Accuracy

Accuracy denotes to the ability of the model to correctly predict the class label of new or unseen data. It is calculated as-

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 1. Accuracy

ITERATION	EXISTING	PROPOSED
1	91.9271	95.4479
2	91.9271	95.3177
3	97.0052	99.3958
4	97.3958	99.9648
5	98.4979	99.8590
6	98.5979	99.9593
7	98.9583	99.9794
8	97.2188	99.6592
9	97.2188	99.8698
10	97.2188	99.8698

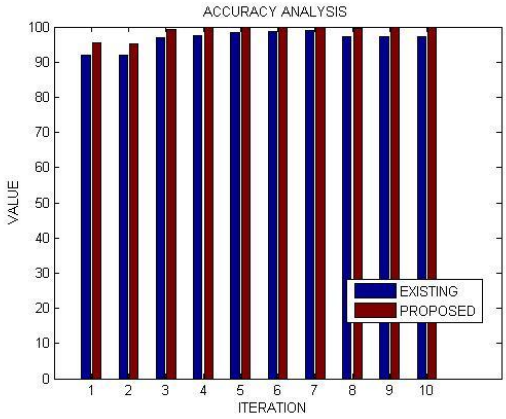

Figure 2. Accuracy Evaluation

B. Mean square error

Mean square error is a measure of image quality index. The large value of mean square ensures that image is a poor quality. Mean square error between the reference image and the fused image is :

$$MSE = \frac{1}{ab}\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - Y_{ij})^2$$

Where Aij and Bij are the image pixel value of reference image.

Table 2. Mean Square Error

ITERATION	EXISTING	PROPOSED
1	0.0807	0.0755
2	0.0788	0.0723
3	0.079	0.0725
4	0.0765	0.0739
5	0.0771	0.0732
6	0.0773	0.0738
7	0.077	0.0742
8	0.0778	0.0739
9	0.0786	0.0729
10	0.079	0.0721

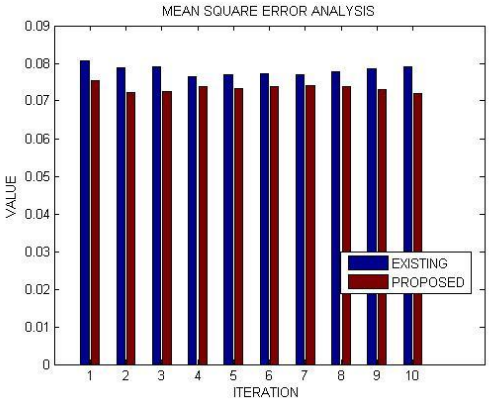

Figure 3. Mean Square Error Evaluation

C. Sensitivity

Sensitivity measures the proportions of positive that are correctly identified. It lies between 0-1. Value of sensitivity near to 1 signifies efficient results.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Table.3. Sensitivity

ITERATION	EXISTING	PROPOSED
1	0.919	0.974
2	0.919	0.983
3	0.97	0.984
4	0.974	0.997
5	0.987	0.993
6	0.987	0.996
7	0.99	0.996
8	0.992	0.995
9	0.992	0.997
10	0.992	0.999

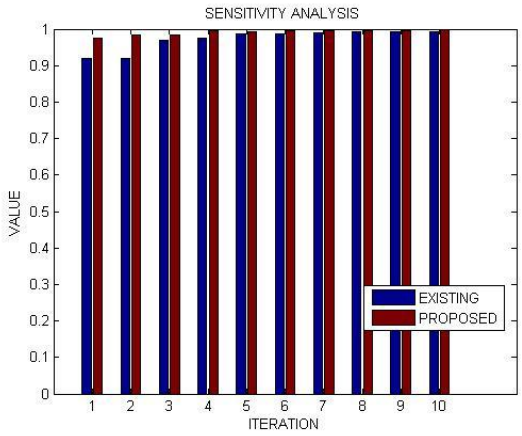

Figure 4. Sensitivity Evaluation

VI. CONCLUSION AND FUTURE SCOPE

Data mining is the computational procedure for discovering routines within big files portions ("big files") pertaining to techniques in the intersection involving synthetic contemplating capability, unit learning, data, as well as collection programs. In this paper, we have proposed a new method in order to improve the accuracy of diabetes classification rate. The proposed technique have integrated Particle swarm optimization (PSO) with support vector machine (SVM) based machine learning technique. The proposed technique also verified by using the various standard diabetes classification data sets. The comparison drawn among the proposed and the existing technique based upon the various standard quality metrics of the data mining. Experimental results indicate that the proposed algorithm is more efficient than existing techniques.

ACKNOWLEDGMENT

I Would like to thank Mr. Prabhdeep Singh and Khalsa College of Engineering and Technology for their support and guidance.

REFERENCES

[1] Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., & Hammond, W.E.: Medical data mining: knowledge discovery in a clinical data warehouse. In Proceedings of the AMIA annual fall symposium (p. 101). American Medical Informatics Association(1997).

[2] Parpinelli, R. S., Lopes, H. S., &Freitas, A. A.: An ant colony based system for data mining: applications to medical data. In Proceedings of the genetic and evolutionary computation conference (GECCO-2001) (pp. 791-797)(2001,July).

[3] Ghazavi, S. N., & Liao, T. W.: Medical data mining by fuzzy modeling with selected features. Artificial Intelligence in Medicine, 43(3), 195-206(2008).

[4] Delen, D., Walker, G., &Kadam, A.:Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127(2005).

[5] Moses, D.:A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data. Kuwait Journal of Science, 42(2)(2015).

[6] Ji, S., Wang, Z., Liu, Q., & Liu, X.:Classification Algorithms for Privacy Preserving in Data Mining: A Survey. In International Conference on Computer Science and its Applications (pp. 312-322). Springer Singapore(2016,December).

[7] Rani, G., Gladis, D., &Mammen, J.Classification and Prediction of Breast Cancer Data derived Using Natural Language Processing. In Proceedings of the Third International Symposium on Women in Computing and Informatics (pp. 250-255). ACM(2015,August).

[8] Das, T. K.:A customer classification prediction model based on machine learning techniques. In 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (pp. 321-326). IEEE(2015,October).

[9] Yu Zhou, Xiaokang Yang, Yongzheng Zhang, Xiang Xu, Yipeng Wang, Xiujuan Chai, Weiyao Lin, Unsupervised adaptive sign language recognition based on hypothesis comparison guided cross validation and linguistic prior filtering, Neurocomputing, Volume 149, Part C, 3 February 2015, Pages 1604-1612.

[10] WalidMagdy, Tamer Elsayed, Unsupervised adaptive microblog filtering for broad dynamic topics, Information Processing & Management, Volume 52, Issue 4, July 2016, Pages 513-528.

[11] Shiqiang Du, Yide Ma, Shouliang Li, Yurun Ma, Robust unsupervised feature selection via matrix factorization, Neurocomputing, Volume 241, 7 June 2017, Pages 115-127.

[12] Daniel Carlos GuimarãesPedronette, Ricardo da S. Torres, Unsupervised Rank Diffusion for Content-Based Image Retrieval, Neurocomputing, Available online 16 May 2017.

[13] Herman Kamper, Aren Jansen, Sharon Goldwater, A segmental framework for fully-unsupervised large-vocabulary speech recognition, Computer Speech & Language, Available online 18 May 2017.

[14] Wei He, Xiaofeng Zhu, Debo Cheng, Rongyao Hu, Shichao Zhang, Unsupervised feature selection for visual classification via feature-representation property, Neuro-computing, Volume 236, 2 May 2017, Pages 5-13.

[15] ZouhairMbarki, HasseneSeddik, Ezzedine Ben Braiek, A rapid hybrid algorithm for image restoration combining parametric Wiener filtering and wave atom transform, Journal of Visual Communication and Image Representation, Volume 40, Part B, October 2016, Pages 694-707.

[16] Gloria Re Calegari, EmanuelaCarlino, Diego Peroni, Irene Celino, Filtering and windowing mobile traffic time series for territorial land use classification, Computer Communications, Volume 95, 1 December 2016, Pages 15-28.

[17] Mostafa Mohammadpourfard, Ashkan Sami, AlirezaSeifi, A statistical unsupervised method against false data injection attacks: A visualization-based approach, Expert Systems with Applications, Volume 84, 30 October 2017, Pages 242-261.

[18] EliahuKhalastchi, Meir Kalech, LiorRokach, A hybrid approach for improving unsupervised fault detection for robotic systems, Expert Systems with Applications, Volume 81, 15 September 2017, Pages 372-383.

[19] Chong Yang, Xiaohui Yu, Yang Liu, YanpingNie, Yuanhong Wang, Collaborative filtering with weighted opinion aspects, Neurocomputing, Volume 210, 19 October 2016, Pages 185-196.

[20] Pengfei Zhu, Wencheng Zhu, Qinghua Hu, Changqing Zhang, WangmengZuo, Subspace clustering guided unsupervised feature selection, Pattern Recognition, Volume 66, June 2017, Pages 364-374.

Authors Profile

Ramandeep kaur Completed B.TECH in Information Technology branch from ACET Amritsar In 2014. She is now PURSUING M.TECH Degree in Computer science and Engineering branch from kcet Amritsar.

Er.Prabhdeep Singh Completed M.TECH in Computer Science and Engineering branch from THAPAR University. He is now pursuing PHD from Punjabi University. His research areas are Machine learning, Cloud Computing and Fog Computing. He is Published various papers on renowned journal.

