Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 4, April 2016, pg.144 – 147

Classification of Imbalanced Medical Data Efficiently Using Multiclass SVM and Genetic Algorithm

Ankit R. Deshmukh[1], Prof. S. P. Akarte[2]

[1]M.E. (CSE), Second Year, Dept. of Computer Science, Prof. Ram Meghe Institute of Technology and Research, Badnera, Amravati, Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701

[2]Assistant Professor, Dept. of Computer Science, Prof. Ram Meghe Institute of Technology and Research, Badnera, Amravati, Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701

[1] ankitdeshmukh100@gmail.com; [2] s_akarte25@rediffmail.com

Abstract—We focused on developing efficient support vector machine based on genetic algorithm for multiclass classification from large of collection imbalanced data. And gives well sorted data. Many years ago, the scientific community is concerned about how to increase the accuracy of different classification methods, and major achievements have been made so far. Hadoop and MapReduce are used to handle these large volumes of variable size data. In proposed system we can take sufficient .csv file as inputs & we apply variance algorithm & generate expected results. Classification is method to perform on poorly & minority class examples when the dataset is extremely imbalanced.

Keywords: Classification, Genetic Algorithm, Hadoop, Map-Reduce, Support Vector Machines.

I. INTRODUCTION

A well balanced dataset is very important for creating a good prediction model. Medical datasets are often not balanced in their class labels. Most existing classification methods tend to perform poorly on minority class examples when the dataset is extremely imbalanced. This is because they aim to optimize the overall accuracy without considering the relative distribution of each class [1]. In the field of biomedical, the issue of learning from these imbalanced data is highly important because it can invent useful knowledge to make important decision on the other hand it can also be extremely costly to misclassify these data. In machine learning, multiclass or multinomial classification is the problem of classifying instances into more than two classes [2]. While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies.

Classification for imbalance data sets has been studied by machine learning community since the last decade [4]. The under sampling and over sampling method [5] balances the data sets by randomly selecting small number of objects from majority class, and doubling the objects in the minority class.The main drawback is some import points, such as support vectors, may be neglected by the random algorithm [3] pointed out the under sampling strategy is not a good choice for SVM, and the over sampling cannot improve the final accuracy.

II. LITERATURE REVIEW

Imbalanced data is a common and serious problem in many biomedical classification tasks. It creates lots of confusion on the training of classifiers and results in lower accuracy of minority classes prediction [6]. This is engendered due to the less availability or by limitations on data collection process such as high cost or privacy problems. The majority classes' overloads standard machine learning algorithms that it cannot bear the load and traditional classifiers making confusions between the decisions towards the majority class and try to optimize overall accuracy. To improve traditional methods or to develop new algorithms for solving the problem of class imbalance, many researches were done. Most of those studies are focused only on binary case or two classes. Only a few researches have been done for multiclass imbalance problem which are more common and complicated in structure in the real-world application.

III. HADOOP OVERVIEW

Hadoop is a distributed computing framework released by Apache Foundation, it is Google's open source implementation of the cloud computing model, and it can be efficient, reliable, scalable way to process data. Its core idea is to build on a large number of cheap and efficient cluster hardware devices, in the form of software processing to provide storage and computing environment for the huge amounts of data, and provide a unified standard interface, is a highly scalable distributed computing systems. Hadoop is a Map Reduce programming model and mass data. It has made a lot of simulation system in the cloud computing, such a calculation based on the concept of cloud modeling and simulation platform of COSIM-CSP system, a new mode of the networked manufacturing, private cloud framework for visual simulation, and the military training system.

IV. HADOOP MAPREDUCE

The term Hadoop [7] comprises a family of many related projects with the same infrastructure for distributed computing and large-scale data processing. It is better known for the Map Reduce algorithm, shown below, and its distributed file system HDFS, which runs on large clusters of commodity machines. Hadoop was created by Doug Cutting and has its origins in Apache Nuts, an open source web search engine. In January 2008 Hadoop was made a top-level project at Apache, attracting to itself a large active community, including Yahoo!, Facebook and The New York Times. At present, Hadoop is a solid and valid presence in the world of cloud computing. Map Reduce is a programming model whose origins lie in the old functional programming. It was adapted by Google as a system for building search indexes, distributed computing and database communities. It was written in C++ language and was made as a framework, in order to simply develop its applications. In Hadoop programs are mainly in Java language but it is also possible, through a mechanism called "streaming", to develop programs in any language that supports the standard I/O. Map Reduce is a batch query processor and the entire dataset is processed for each query. It is a linearly scalable programming model where users programs at least two functions: the "map" function and "reduction" functions. These functions process the data in terms of key/value pairs which are unaware of the size of the data or the cloud that they are operating on, so they can be used unchanged either for a small dataset or for a massive one.

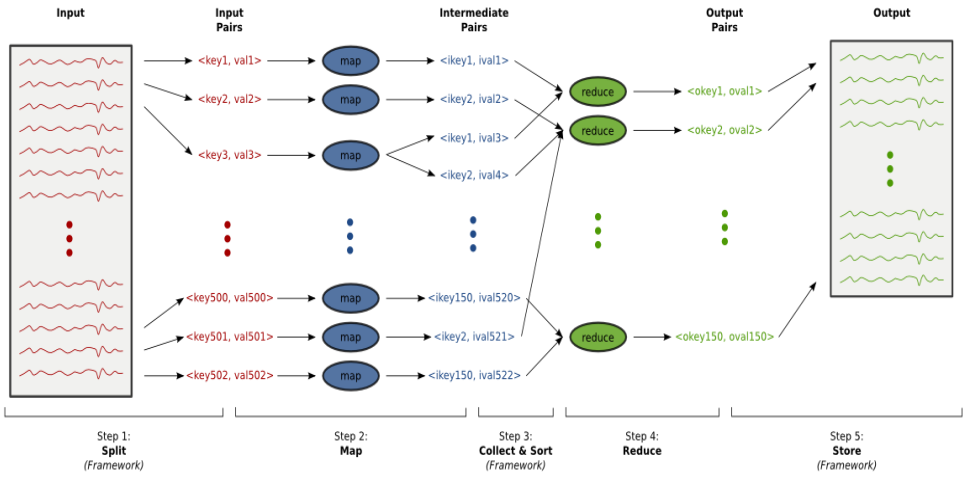

Fig. 1 Mapping & Reduction

V. GENETIC ALGORITHM

The Genetic Algorithm (GA) is an optimization and search technique based on the principles of genetics and natural selection. Generally GAs are not used to find patterns, but rather to guide the learning process of data mining algorithms such as neural nets. A GA allows a population composed of many individuals [8] (basically the candidates) to evolve under specified selection rules to a state that maximizes the fitness. GA is known as a subset of evolutionary algorithms that model biological processes which is influenced by the environmental factor to solve various numerical optimization problems. GA allows a population composed of many individuals or called chromosomes to evolve under specified rules to a state that maximizes the fitness or minimizes the cost functions. A genetic algorithm mainly composed of three operators: selection, crossover, and mutation. In selection, a good string (on the basis of fitness) is selected to breed a new generation; crossover combines good strings to generate better offspring; mutation alters a string locally to maintain genetic diversity from one generation of a population of chromosomes to the next. In each generation, the population is evaluated and tested for termination of the algorithm. If the termination criterion is not satisfied, the population is operated upon by the three GA operators and then re-evaluated. The GA cycle continues until the termination criterion is reached. In feature selection, Genetic Algorithm (GA) is used as a random selection algorithm, Capable of effectively exploring large search spaces[9].

A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a metaheuristic) is routinely used to generate useful solutions to optimization and search problems [10].

VI. SUPPORT VECTOR MACHINE

A. Title and Author Details

Support Vector Machine (SVM) is inspired on statistical learning theory developed by Vapnik on 70's [13]. It achieves optimal classification in linear separable case. It is better than neural networks [14], decision trees [15] and Bayesian classifiers [16] in some applications. SVM offers a hyperplane that represents the largest separation (or margin) between two classes [11]. This kind of maximum-margin hyperplane may not exist because of class overlapping or mislabeled examples. The soft margins SVM by introducing slack variables can find a hyperplane that splits the examples as cleanly as possible. However, SVM requires balance data and it does not consider the classes' distribution. Support Vector Machine (SVM) is a classification technique based on statistical learning theory. It is based on the idea of a hyper plane classifier. The goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized [12].

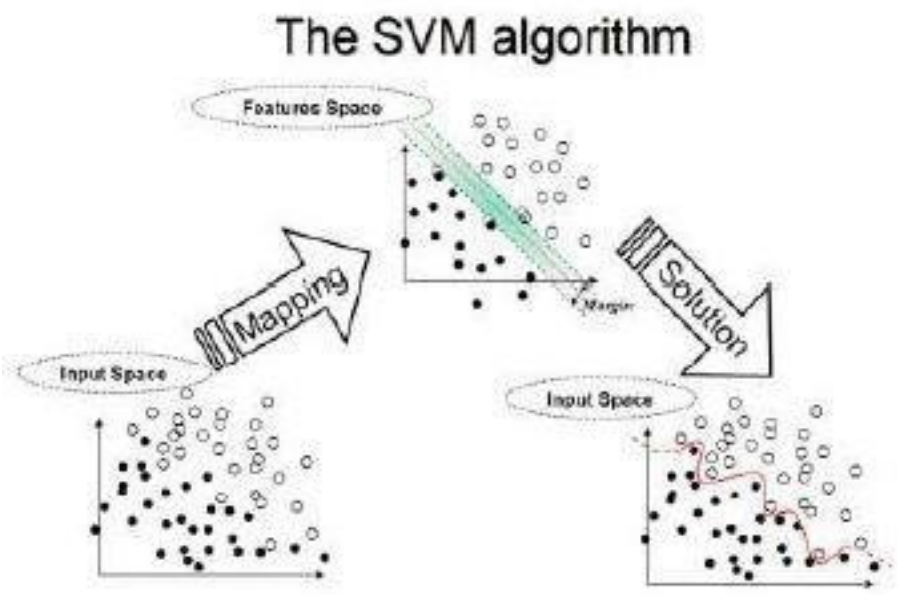

Fig. 2 SVM Process

VII. CONCLUSION

The proposed system uses the support vector machine and genetic algorithm to works & processes the multiclass data sets. And doing so it is found that the algorithm works correctly and this multiclass classification is beneficial for saving time required for classification process on large scale, also for saving the storage space on storage space by avoiding repetition of data.

REFERENCES

[1] Chawla N V, Japkowicz N. Editorial: Special issue on learning from imbalanced datasets. SIGKDD Explorations, 2004, 6: 1-6.
[2] R. Laza *et al.*, "Evaluating the effect of unbalanced data in biomedical document classification," *Journal of integrative bioinformatics*, vol. 8, no. 3, pp. 177, 2011 Sep,2011.
[3] Suzan Koknar-tezel and Longin Jan Latecki,Improving SVM Classi_cation on Imbalanced Data Sets in Distance Spaces, *IEEE International Conference on Data Mining*, 2009, pp 259 - 267.
[4] Z.-Q. Zeng and J. Gao, "Improving svm classi_cation with imbalance data set," in *Proceedings of the 16th International Conference on Neural Information Processing*, Springer- Verlag, 2009, pp.389–398.
[5] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," *In Proc. ECML*, 2004, pp.39-50.
[6] Chawla N V, Japkowicz N. Editorial: Special issue on learning from imbalanced datasets. SIGKDD Explorations, 2004,6: 1-6.
[7] T. White, *Hadoop: The Definitive Guide*, Third. O'Reiily, 2012.
[8]. Jihoon Yang and Vasant Honavar. Feature subset selection using Genetic Algorithm. IEEE Intelligent Systems, 1998.
[9]. L. Chu, and C. Wu, "A Fuzzy Support Vector Machine Based on Geometric Model," Proceedings of the fifth World Congress on Intelligent Control and Automation, Hangzhou, P.R. China, pp.1843-1846, June 15-19, 2004.
[10] Mitchell, Melanie (1996). An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press. ISBN 9780585030944.
[11] N.Cristianini, J.Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* , Cambridge University Press, 2000.
[12] C.J.C. Burges, " A tutorial on support vector machines for pattern recognitation " , Data Mining and Knowledge Discovery , vol.2 , 1998, pp. 121-167.
[13] Vapnik V., "The Nature of Statistical Learning Theory," *Springer, N.Y.*, 1995.
[14] Ra¸sit Köker, A genetic algorithm approach to a neuralnetworkbased inverse kinematics solution of robotic manipulators based on error minimization, Information Sciences, Volume 222, 10 February 2013, Pages 528-543.
[15] J. Chen, C. Wang, and R. Wang, Combining support vector machines with a pairwise decision tree, *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 409 –413, july 2008.
[16] J. Cervantes, X. Li, and W. Yu, "Splice site detection in dna sequences using a fast classi_cation algorithm," 2009 *IEEE international conference on Systems, Man and Cybernetics*, Piscataway, NJ, USA, 2009, pp. 2683–2688.

