Classification Mining SNPs from Leukaemia Cancer Dataset Using Linear Classifier with ACO

P.H. Prathibha and Dr.C.P. Chandran

Abstract--- Single Nucleotide Polymorphisms(SNP) are the foremost common type of genetic variation in human comprising nearly 1/1,000th of the typical human genome. SNP offer the foremost complete information for genome-wide association studies. We tend to propose a process methodology to quickly notice true SNPs in public-available leukemia cancer database. Much analysis has been specializing in various genetic models to spot genes that may predict the disease status. However, increasing the amount of SNPs generates large amount of combined genetic outcomes to be tested. Classification could be a data processing technique to predict cluster membership for data instances. The ACO could be a probabilistic technique for computational issues which may be reduced to finding sensible ways through graphs. In this research paper data mining classification techniques linear classifier are analyzed with ACO on leukemia cancer dataset. Performance of these techniques is compared by accuracy, Sensitivity and Specificity. The experimental results show that Linear classifier with ACO is able to distinguish cancer diseases from normal with the maximum accuracy of 73.20%, Sensitivity of 69.21% and specificity of 65% whereas SVM are 70.00% of accuracy, 65.20% of sensitivity and specificity of 63.53%.

Keywords--- SNP, Single Nucleotide Polymorphism, Data Mining, ACO, Classification, Linear Classifier

I. INTRODUCTION

Data Mining

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [1]. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness, metrics, complexity considerations, post processing of discovered, structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD [2].

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records, unusual records and dependencies. This usually involves using database techniques. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and . For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a [3].

Classification

Classification is that the downside of characteristic to that of a collection of classes a replacement observation belongs.on the idea of a coaching set of knowledge containing observations whose class membership is thought associate example would be assignment a given email into "spam" or "non-spam" categories or assignment a identification to a given patient as delineated by discovered characteristics of the patient[4]. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category [5].

Linear Classifier

In the field of machine learning, the goal of applied math category classification is to use Associate in identify object's characteristics to spot that class it belongs to.A linear classifier achieves this by creating a classification call supported the worth of a linear combination of the characteristics [6]. Associate in identifying object’s characteristics are referred to as feature values and are generally conferred to the machine during a vector referred to as a feature vector. Such classifiers work well for sensible issues like document classification, and a lot of use for issues with several variables reaching accuracy levels adore non-linear classifiers whereas taking less time to coach and use [7]. Linear models for classification separate input vectors into categories victimization linear call boundaries. For Classification issues, we wish y to be a predictor of t. In different words, we have a tendency to want to map the input vector into one amongst variety of distinct categories, or to posterior chances that lie between zero and one. For this purpose, it's helpful to elaborate the linear model...
by introducing a nonlinear activation function, which usually can constrain y to lie between -1 and one or between zero and one. A classifier based mostly upon this straightforward generalized linear model is termed a perceptron. It also can be known with associate degree abstracted model of a neuron cell known as the McCulloch Pitts model.

For an important class of procedures, these decision boundaries are linear. This is what we will refer to as linear methods for classification. We will see that these can be quite flexible (much more than linear regression). Suppose we have K classes labeled 1, . . . , K and a single predictor X. We can define a 0–1 indicator for each class k, and perform regression for each of these. We would end up with a regression function \( f_k(x) = \beta_0^k + \beta_1^k x \) for each class k. The decision boundary between class k and l is simply the set for which \( f_k(x) = f_l(x) \), i.e., \( \{x : \beta_0^k + \beta_1^k x = \beta_0^l + \beta_1^l x\} \) which is a hyper plane. Since the same holds for any pair of classes, the division of the input space is piecewise linear. [8].

Genomics

Genomics also contains the genes and their function. Genomics aims to understand the structure of the genome, including the mapping genes and sequencing the DNA. Genomics examines the molecular mechanisms and the interplay of genetic and environmental factors in disease. Genomics includes the four types. They are 1) Functional genomics 2) Structural genomics 3) Comparative genomics 4) Epigenomics. The Functional genomics are used to characterization of genes and their mRNA and protein products. Structural genomics are used to dissection of the architectural features of genes and chromosomes. Comparative genomics are used to the evolutionary relationships between the genes and proteins of different species. Epigenomics are used to the DNA methylation patterns, imprinting and DNA packaging. Pharmacogenomics are used to new biological targets and new ways to design drugs and vaccines [9].

Single Nucleotide Polymorphism (SNP)

A single nucleotide polymorphism or simple nucleotide polymorphism, often abbreviated to just , is a variation in a single nucleotide which may occur at some specific position in the genome, where each variation is present to some appreciable degree within a population (e.g. >1%) [10].

For example, at a specific base position in the human genome, it may be that in most individuals the base C appears there; but in a minority of individuals, the base A appears at that position instead [11]. There is an SNP at this specific base position, and the two possible nucleotide variations - C or A - are said to be alleles for this base position. Although in this example and most SNPs so far discovered there are only two different alleles, there are also triallelic SNPs in which three different base variations may coexist within a population [12].

SNPs underlie differences in our susceptibility to disease; a wide range of human diseases, e.g. sickle-cell anemia, \( \beta \)-thalassemia and cystic fibrosis result from SNPs. The severity of illness and the way our body responds to treatments are also manifestations of genetic variations. For example, a single base mutation in the APoE gene is associated with a higher risk for Alzheimer's disease [12].

\[ \text{Fig. 1: DNA Molecule 1 Differs from DNA Molecule 2 at a Single Base-Pair Location} \]

Ant Colony Optimization (ACO)

In computer science and operations research, the ant colony optimization algorithm (ACO) is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs [13].

This algorithm is a member of the ant colony algorithms family, in swarm intelligence methods, and it constitutes some metaheuristic optimizations. Initially proposed by Marco Dorigo in 1992, the first algorithm was aiming to search for an optimal path in a graph, based on the behavior of ants seeking a path between their colony and a source of food. The original idea has since diversified to solve a wider class of numerical problems, and as a result, several problems have emerged, drawing on various aspects of the behavior of ants [14].

The ant colony algorithm is an algorithm for finding optimal paths that is based on the behavior of ants searching for food. At first, the ants wander randomly. When an ant finds a source of food, it walks back to the colony leaving "markers" that show the path has food. When other ants come across the markers, they are likely to follow the path with a certain probability. If they do, they then populate the path with their own markers as they bring the food back. As more ants find the path, it gets stronger until there are a couple streams of ants traveling to various food sources near the colony. Because the ants drop pheromones every time they bring food, shorter paths are more likely to be stronger, hence optimizing the "solution." In the meantime, some ants are still randomly scouting for closer food sources. A similar approach can be used find near-optimal solution to the traveling salesman problem. Once the food source is depleted, the route is no longer populated with pheromones and slowly decays. Because the ant-colony works on a very dynamic system, the ant colony algorithm works very well in graphs with changing topologies. Examples of such systems include computer networks, and artificial intelligence simulations of workers [15].

II. PREVIOUS WORKS

Picoult-Newberg et al.; (1999) first proposed mining SNPs from EST sequences. They selected SNPs from candidates using four filters. The four filters can eliminate false alignments, gaps and base-calling errors [16]. Taillon-Miller et al.; (1998) did the same using overlapping genomic sequences. At the same time, Marth et al.; (1999) developed POLYBAYES for SNPs discovery, which was based on a simple-machine learning method: Naive Bayes. After that, many researchers looked for ways to distinguish SNPs from
sequence errors [17]. Besides the filters, Batley et al.; (2003) proposed two scores, which helped them find 264 SNP candidates in maize. Softwares and web servers were also developed for mining SNPs from data, including autoSNP (Barker et al.;, 2003), SEAN (Huntley et al; 2006), SNPServer (Savage et al.;, 2005), and SNP detector (Zhang et al.;, 2005) [18,19].

SNPs are delicate variations, like insertions, deletions and substitutions ascertained within the genomic DNA sequences of people of a similar species (2006). It's distinguished from rare variations by a demand for the smallest amount abundant cistron to possess a frequency of one hundred and twenty fifth or additional. There are many blessings of mistreatment SNP information over microarray information [20]. First, SNPs are unlikely to alter over time. That is, SNP pattern of a patient is probably going to be as similar as that collected at one's birth or early in life (2008). Second, it will be collected from any tissue within the body so it's comparatively simple compared to the microarray that should be taken from the neoplasm tissue

Lu et al.; [22] proposed Fast Genetic K-means cluster technique (FGKA). It is a faster version of GKA and FGKA that features several improvements over GKA including an efficient evaluation of the objective value TWCV (Total Within-Cluster Variation), avoiding illegal string elimination overhead, and a simplification of the mutation operator. Mohammad Ali proposed [23] a hybrid genetic k-means algorithm (HGKA).


III. METHODOLOGY USED

The methodology employed in this work is shown within the Fig2. The input is Leukemia Dataset. Linear Classifier with ACO algorithmic program is employed to predict the affected disease. The SNP candidate set consists of true SNPs and pseudo-SNPs. The mining formula must distinguish between SNPs from the pseudo-SNPs. True SNP identification will be remodeled into a classification downside. Between SNPs correspond to positive samples and also the pseudo-SNPs correspond to negative samples. The mining formula must construct a classifier to partition the positive samples and also the negative samples. There essential steps within the classifying procedure; one is that the classifying feature choice and also the different is that the Linear classifier with ACO.

Fig. 2: Classification Mining of SNP

Our aim is to increase the level of performance classification techniques nowadays by using optimization method like ACO. In this Linear classifier with ACO is used to increase the accuracy for better results. We choose Linear Classifier with ACO in this work. Here in this paper we can measure the performance based on train values of these methods in the Leukaemia data set.

IV. PROPOSED ALGORITHM

Classification is a supervised learning and is one of the most studied data mining technique. In linear classification, the input feature vector to the classifier is a real vector, then the output score is where a real vector of weights is and f is a function that converts the dot product of the two vectors into the desired output. In other words, it's a one-form or linear functional mapping onto R.

The weight vector is learned from a set of labeled training samples. Often f is a simple function that maps all values above a certain threshold to the first class and all other values to the second class. A more complex might give the probability that an item belongs to a certain class.

For a two-class classification problem, one can visualize the operation of a linear classifier as splitting a high-dimensional input space with a hyperplane: all points on one side of the hyperplane are classified as "yes", while the others are classified as "no". A linear classifier is often used in situations where the speed of classification is an issue, since it is often the fastest classifier, especially when sparse. Also, linear classifiers often work very well when the number of dimensions in is large, as in document classification, where each element in is typically the number of occurrences of a word in a document. In such cases, the classifier should be well-regularized. Fig. 3 shows the Linear Distance Measures.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

![Distance functions](image)

Fig. 3: Distance Measures for Linear
It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset. [26].

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Given a training set D and a test object x = (x′, y′), the algorithm computes the distance (or similarity) between z and all the training objects (x, y) ∈ D to determine its nearest-neighbor list, Dz. (x is the data of a training object, while y is its class. Likewise, x′ is the data of the test object and y′ is its class.) Once the nearest-neighbor list is obtained, the test object is classified based on the majority class of its nearest neighbours.

The ant colony optimization algorithm (ACO) is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. This algorithm is a member of the ant colony algorithms family, in swarm intelligence methods, and it constitutes some metaheuristic optimizations. In general, ACO algorithms employ a finite size of artificial agents with defined characteristics which collectively search for good quality solutions to the problem under consideration. Starting from an initial state selected according to some case-dependent criteria, each ant builds a solution which is similar to a chromosome in a genetic algorithm.

While building its own solution, each ant collects information on its own performance and uses this information to modify the representation of the problem, as seen by the other ants. The ant’s internal states store information about the ant’s past behavior, which can be employed to compute the goodness/value of the generated solution.

Let τij(t) be the pheromone deposited on path ij at time t, and ηij(t) be the heuristic value of path ij at time t according to the measure of the objective function. Defining the transition probability from node i to node j at time period t as [27].

$$P_{ij}(t) = \left\{ \begin{array}{ll} \frac{[\eta_{ij}(t)]^{\alpha}[\tau_{ij}(t)]^{\beta}}{\sum [\eta_{ij}(t)]^{\alpha}[\tau_{ij}(t)]^{\beta}} & \text{if } j \in \text{allowed} \\
\text{otherwise} & \end{array} \right.$$  

(1)

where α and β = parameters that control the relative importance of the pheromone trail versus a heuristic value. Let q be a random variable uniformly distributed over [0, 1], and q0 ∈ [0, 1] be a tunable parameter. The next node j that ant k chooses to go is [28].

$$j = \arg \max \left\{ \left[ \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum [\eta_{ij}(t)]^{\alpha}[\tau_{ij}(t)]^{\beta}} \right] \right\} \text{if } q \leq q_0$$

otherwise (2)

where J = a random variable selected according to the probability distribution of Pij(t) (See Eq. (1)). The pheromone trail is changed globally. Upon completion of a tour by all ants in the colony, the global trail updating is done as follows:

$$\tau_{ij}(t) = (1-\rho) \cdot \tau_{ij}(t) + \rho \cdot \Delta \tau_{ij}$$

(3)

where 0 ≤ ρ ≤ 1; (1 - ρ) = evaporation rate; and the symbol ρ iteration is used to show the next iteration. Main ACO algorithms are Ant System (AS), Max-Min Ant System (MMAS), and Ant Colony System (ACS). This paper, we have used Ant Colony System (ACS). Fig.4. Shows the proposed Linear classifier with ACO algorithm. Fig5. Shows the flowchart depicting the Linear classifier with ACO Algorithm.

1) Initialization
- Determine the population of ants
- Set the initial pheromone intensities for each ant
- Set the ACO parameters such as α, β, λ and max-iterations

// Linear Classification Algorithm
Read the training data from a file
Read the testing data from a file
Set K to some value
Normalize the attribute values in the range 0 to 1
Vote = Value / (1+Value);
Apply Linear Classifier
sign(ax + by + c)
where a, b, and c are constants that define a line.
This separating line can be found by looking at the line
ax + by + c = 0
Apply Backward Elimination
For each testing example in the testing data set
2) while (gene found) do
I ← All genes in E
J ← All conditions in E
Addition Phase(LJ)
DeletionPhase(LJ)
Report genes LJ
Mask(LJ)
3) Evaluation of ants selected subsets
4) Check the stop criterion
5) Go to step 8, if the number of iterations is more than the max-iterations, otherwise continue.
6) Pheromone updating
7) Generation of new ants and create the new feature subsets
8) In this step previous ants are removed and new ants are generated. Each generated ant chooses a path according to probability rule.
9) Go to step 3 and continue.
10) Evaluation the classification accuracy of final subset
11) For final evaluation of the selected subset, an ANN is used as classifier.

Fig.4: Proposed Linear Classification with ACO Algorithm

During initialization, each Bimax Bk (where 1 ≤ k ≤ K ) is initialized to bimax BA (the location of the home colony). BA contains all the rows and columns of the original matrix.
V. DATASET USED

Leukaemia could be a cluster of cancers that sometimes begins within the bone marrow and leads to high numbers of abnormal white blood cells. These white blood cells don't seem to be absolutely developed and area unit referred to as blasts or cancer of the blood cells. Symptoms might bleed and bruising issues, feeling terribly tired, fever and an enhanced risk of infections. These symptoms occur due to an absence of normal blood cells diagnosing is often by blood tests or bone marrow biopsy [29].

The exact cause of leukemia is unknown. Different kinds of leukemia are believed to have different causes. Both inherited and environmental factors are believed to be involved. Risk factors include smoking, ionizing radiation, some chemicals, prior chemotherapy, and Down syndrome. People with a family history of leukemia are also at higher risk. There are four main types of leukemia: acute lymphoblastic leukemia, acute myeloid leukemia chronic lymphocytic leukemia and chronic myeloid leukemia, as well as a number of less common types. Leukemia is part of a broader group of neoplasms which affect the blood, bone marrow, and lymphoid system, known as tumors of the hematopoietic and lymphoid tissues [30]. Table 1 shows the information of 28 SNPs database.

Table 1: Information of 28 SNPs database of Leukaemia Cancer Patient

<table>
<thead>
<tr>
<th>Grp</th>
<th>No</th>
<th>SNPs</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>CCR5(-459)</td>
<td>G/A</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>RANTES(-403)</td>
<td>G/A</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>MCP1(-2518)</td>
<td>G/A</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>CCR2(-64)</td>
<td>G/A</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>CXCR1(-2761)</td>
<td>G/A</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>CXCR4(-1381)</td>
<td>G/A</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>IL1B(-31)</td>
<td>C/T</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>IL1B(-31)</td>
<td>C/T</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>ILIRN(-1308)</td>
<td>C/T</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>IL1RN(+3UTR)</td>
<td>C/G</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>MBP(-543)</td>
<td>A/G</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>IFIRF(-410)</td>
<td>G/A</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>IFNGR2(-64)</td>
<td>G/A</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>IFIRF(-388)</td>
<td>C/T</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>IL(-102)</td>
<td>A/C</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>IL(-1082)</td>
<td>G/A</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>IFIRF(-56)</td>
<td>C/T</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>IFIRF(+95)</td>
<td>C/T</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>IFIRF(+874)</td>
<td>A/T</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>TNF(-238)</td>
<td>G/A</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>TNF(-308)</td>
<td>G/A</td>
</tr>
<tr>
<td>22</td>
<td>22</td>
<td>IL18(-535)</td>
<td>G/A</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>MMP3(-455)</td>
<td>G/A</td>
</tr>
<tr>
<td>24</td>
<td>24</td>
<td>MMP3(-96D)</td>
<td>G/A</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>MMP3(-362A)</td>
<td>C/T</td>
</tr>
<tr>
<td>26</td>
<td>26</td>
<td>MMP9(-279Q)</td>
<td>G/A</td>
</tr>
<tr>
<td>27</td>
<td>27</td>
<td>MMP9(-688R)</td>
<td>G/A</td>
</tr>
<tr>
<td>28</td>
<td>28</td>
<td>MMP9(-697G)</td>
<td>C/A</td>
</tr>
<tr>
<td>29</td>
<td>29</td>
<td>MMP3(-96D)</td>
<td>C/A</td>
</tr>
</tbody>
</table>

VI. RESULTS AND DISCUSSION

The experimental results are compared based on the following metrics. Accuracy of a test can be expressed through sensitivity and specificity. Sensitivity refers to the ability of a certain diagnostic test to detect a cancer. It is expressed as the probability of testing positive if the particular disease is truly present, i.e., the probability of having both positive test and positive diagnosis. Therefore, a test with 90% sensitivity means that 90% of those with cancer will test positive. Specificity, on the other hand, refers to the probability of testing negative if chronic hepatitis is truly absent. Another parameter, accuracy is the ratio of the total of truly positive and truly negative over the total number of patients. Table 2 shows the calculation of sensitivity, specificity and accuracy.

Table 2: Shows the Calculation of Sensitivity, Specificity and Accuracy

<table>
<thead>
<tr>
<th>Leukemia Cancer</th>
<th>Test Result</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>True position (TP)</td>
<td></td>
<td>TP + FN</td>
<td>FN</td>
</tr>
<tr>
<td>False position (FP)</td>
<td></td>
<td>FP + TN</td>
<td>TN</td>
</tr>
<tr>
<td>False Negative (FN)</td>
<td></td>
<td>TN + FN</td>
<td>TN</td>
</tr>
<tr>
<td>True Negative (TN)</td>
<td></td>
<td>TP + FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{(TP + TN)}{(TP + TN + FP + FN)} \)

Sensitivity = \( \frac{TP}{TP + FN} \)

Specificity = \( \frac{TN}{TN + FP} \)
We present the classification accuracy, sensitivity and specificity for leukemia cancer using Linear classifier with ACO. As seen in Table 3, the best accuracy of 73.20% is obtained using Linear classifier using ACO.

Table 3: The Best Accuracy, Sensitivity & Specificity of Linear Classifier with ACO

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Classifier with ACO</td>
<td>73.20%</td>
<td>69.21%</td>
<td>65%</td>
</tr>
<tr>
<td>KNN</td>
<td>71.2%</td>
<td>67.5%</td>
<td>64.23%</td>
</tr>
<tr>
<td>SVM</td>
<td>70.00%</td>
<td>65.20%</td>
<td>63.53%</td>
</tr>
</tbody>
</table>

VII. CONCLUSION

The proposed linear classifier with ACO is effectively used for locating an affected gene. For these algorithms, the results are positive. However, the performance results aren’t consistent for all three measurements: accuracy, sensitivity and specificity. The identification of the genetic and biochemistry mechanisms of contractile organ fat deposition is incredibly vital for breeding. However, understanding the genetic basis for contractile organ fat deposition is tough as a result of this attribute is probably going controlled by variety of genes.

REFERENCES